Technical University of Munich
Department of Mechanical Engineering
Chair of Ergonomics

Crew Resource Management Training:
Reliability and Validity in Light of Training Transfer

Patrick M. Gontar

Vollständiger Abdruck der von der Fakultät für Maschinenwesen der Technischen Universität München zur Erlangung des akademischen Grades eines Doktor-Ingenieurs (Dr.-Ing.) genehmigten Dissertation

Vorsitzender:

Prof. Dr.-Ing. Boris Lohmann

Prüfer der Dissertation:

1. Prof. Dr. phil. Klaus Bengler
2. Prof. Dr. phil. Dietrich Manzey

Die Dissertation wurde am 21.11.2017 bei der Technischen Universität München eingereicht und durch die Fakultät für Maschinenwesen am 23.03.2018 angenommen.

Abstract

In today's aircraft, pilots are the only agents who can engage in problem-solving or in developing new strategies when severe technical or procedural problems arise. To be able to do so, airline pilots must undergo not only a strict selection process, but also initial and recurrent training. Pilot training, as any training, faces the major challenge of training transfer. That is whether pilots can apply their learned skills and associated knowledge in real situations.

This publication-based doctoral thesis revisits two major factors known to greatly influence training transfer: feedback reliability and content relevance. Two experiments in the area of crew resource management training are dedicated to each factor. Being a precondition for feedback reliability, the first experiment analyzed the interrater reliability of expert raters when assessing pilots' nontechnical skills as they do in pilot training. As results showed that the reliability is not satisfactory, data from a second experiment were used to develop and implement a nonlinear cross-recurrence approach to suggest a reliable alternative to assess the communication behavior of pilots. In consideration of content relevance, a third experiment analyzed whether the assumptions made by procedures and training are currently met during real airline operation. As we found that assumptions are often violated in real operation, we took a closer look at the assumptions made in training. Based on the results of a fourth experiment, which was dedicated to handling unforeseen events, we conclude that the predictability of training context, which is also one of the major assumptions in procedure design, leads to reduced content relevance and validity in pilot training.

The results of the four papers included in this thesis show that current training practice does not feature sufficient interrater reliability in the performance assessment, and it is also lacking in terms of content relevance and validity in specific cases. Most importantly, training and procedures make assumptions that are not reflected in real operation. Taken together, pilot training should be adapted to give greater consideration to the conditions of daily operations. Besides other recommendations, it seems very important to implement unforeseen elements in training to allow pilots to gather experience in handling unforeseen events, as they would occur in reality.

It is obvious that to increase safety further in an already very safe system is a great challenge; nevertheless, this thesis sheds light on some issues that might rather easily be resolved by inexpensive intervention strategies. However, further validation of the results is required from the research community to underpin the conclusions presented here.

Acknowledgment

This thesis originates from my work as a research associate for the Chair of Ergonomics at the Technical University of Munich. I want to express my gratitude to my supervisor Professor Klaus Bengler not only for providing me with the opportunity to conduct research for more than three years but also for his constant support and the often challenging discussions that always aided my development as well as that of the subject matter. Further, I want to thank Professor Boris Lohmann for serving as chair of the examination committee and Professor Dietrich Manzey for joining the committee as an examiner.

Most of this work would not have been possible without the continuous support of several pilots who always had an answer when it came to operational airline questions. Among those I want to mention are Manfred Mueller, Peter Croeniger, and, of course Carsten Schmidt-Moll. For the opportunity to use their flight simulator and their technical support, I want to thank Professor Florian Holzapfel, Ludwig Drees, Chong Wang, and Christoph Krause.

I was also very lucky to have the best colleagues I could imagine. All were a great support, and I am especially grateful to Martin Goetze, Jonas Radlmayr, Anna Feldhuetter, Jonas Schmidtler, Sonja Schneider, Stephanie Cramer, Christian Gold, Christoph Rommerskirchen, and Markus Zimmermann for always being open to discussions and the scientific discourse. In addition, I want to thank all the students who supported me before, during, and after the conducted studies, especially Christoph Hera.

Serving as representative of the doctoral candidates at my alma mater, I want to thank Michael Klimke, Felix Dietrich, Sarah Kersten, Birgit Spielmann, and Po Sang Lam for supporting me in often delicate discussions on all kinds of academic boards.

This work would not have been possible without the support of all my co-authors. I greatly enjoyed our discussions that often led to new ideas and always provoked new thinking processes. A big thanks goes to my mentor Hans-Juergen Hoermann, who always supported me and asked the unpleasant questions that ultimately led to great improvements in my research. Regarding feedback on my manuscripts, I want to thank Ute Fischer for her constant feedback when it came to linguistics and pilot communication. My work would have been very different if I had not met Immanuel Barshi, who became a dear friend, greatly inspired the research I present in Papers #3 and #4, and, despite the nine-hour time difference, always had time for discussions.

I want to thank my friend Andreas Haslbeck, who introduced me not only to the field of human factors, but also to the Chair of Ergonomics about seven years ago.

Finally, I want to express my gratitude to my whole family and especially to Verena, who has supported me since the first day we met.

~ Patrick

Funding & Proofreading

Part of the research presented here was funded by the German Federal Ministry for Economic Affairs and Energy via the Federal Aeronautical Research Program LuFo IV-2 (20V0803).

All papers included in this thesis were proofread by native English speakers prior to publication. This manuscript also underwent proofreading by native English speakers and earlier versions were influenced by discussions and comments from my supervisor Professor Klaus Bengler and my mentor Hans-Juergen Hoermann. The layout of this thesis was inspired by the work of Markus Zimmermann.

Contents

1 Introduction

Flying means controlling. It is only 115 years since the Wright brothers triggered a new era of powered and heavier-than-air flight: the invention and successful flight of the airplane. Following the example of Otto Lilienthal (2003)—described by Wilbur and Orville Wright as their greatest influence—the two brothers constructed and successfully flew several uncontrolled gliders. Several years of engineering and testing were necessary before the Wright brothers were finally prepared to conduct their first controlled and powered flight at Kitty Hawk on December 17, 1903 (Heppenheimer, 2003). It is the controlling part of operating an airplane that allows the pilot to start, fly, and land where desired. Back in the very early 20th century, controlling an airplane was rather a matter of steering and maneuvering the airplane via rudder, elevator, ailerons, and thrust more than anything else—a condition that quickly developed over the years. Considering the human–machine system, the technical subsystem developed quickly with the introduction of turbojet-propelled aircraft in the 1950s. Earlier problems with aircraft structure and engines became less prominent (Helmreich & Foushee, 1993).

Although it was the introduction of new technology that made flying easier, more comfortable, more efficient, and also safer, the focus of development was foremost on the technical subsystem (e.g., M. G. Cooper, Elliott, & Hartzell, 1987) rather than on human factors relating to pilot crews. The uneven advancements in the human–machine system resulted in the relative shift of accident causes. In fact, data from 1959 to 1989 show that the number of accidents involving inappropriate flight crew behavior increased relative to those caused by technical malfunctions (Helmreich & Foushee, 1993; Kayten, 1993). Even now, in the 21st century, as highly sophisticated technical possibilities and technological changes have found their way into aviation, or were even driven by aviation, conducting a safe flight is still about controlling the airplane (Jensen, 2017).

This implies, in contrast to requirements prevalent in 1903, not merely correctly deflecting control surfaces (as they are now driven via flight computers), but rather managing flight-management computers (Sarter & Woods, 1992, 1994), monitoring auto-flight systems (Bjorklund, Alfredson, & Dekker, 2006; Cymek, Jahn, & Manzey, 2016; Muthard & Wickens, 2003; Sarter, Mumaw, & Wickens, 2007), executing procedures and checklists (Degani & Wiener, 1993, 1997; Haslbeck, Kirchner, Schubert, & Bengler, 2014), handling passenger needs (Peterson et al., 2013; Turner, Griffin, & Holland, 2000), coping with tight-scheduled operations (Gontar, Schneider, Schmidt-Moll, Bollin, & Bengler, 2017), and handling complex or ambiguous malfunctions (Orasanu & Fischer, 2014; Orasanu, Martin, & Davison, 2011). The pilot is now faced with challenges, such as increased cognitive demands (Mosier, 2010; Mosier & Fischer, 2010), specifically in terms of information processing and situation awareness (Durso & Alexander, 2010; Endsley, 1995a, 1995b; Vidulich, Wickens, Tsang, & Flach, 2010).

Further aspects of manual flying skill degradation (Arnold, 2015; Haslbeck & Hoermann, 2016; Haslbeck, Kirchner, et al., 2014; Haslbeck, Schubert, Gontar, & Bengler, 2014; Haslbeck et al., 2012) due to lack of practice and training have developed over the years as the pilots' task has shifted from active control to monitoring (Ephrath & Young, 1981; Mosier & Fischer, 2010; Thomas & Rantanen, 2006). Today, pilots also are exposed to long working shifts that can lead to fatigue problems (Caldwell, 2012; Caldwell et al., 2009; Hoermann, Gontar, & Haslbeck, 2015; Rosekind et al., 1994; Weiland et al., 2013). Recent research also indicates that pilots are not very well prepared to make decisions when faced with unforeseen events (Gontar et al., 2018; Gontar, Porstner, Hoermann, & Bengler, 2015). While military and general aviation bring up different challenges, this thesis refers to civil and commercial aviation only.

Over the years, the reliability of the whole civil and commercial aviation system has increased, such that it is nowadays justifiably considered an ultra-safe system (Amalberti & Wioland, 1997; Makins et al., 2016) resulting in an average worldwide accident probability of 1.61×10^{-6} for each flight (International Air Transport Association, 2017). Nevertheless, the increasing number of daily flights led to a total of 65 accidents in 2016, of which 10 were fatal (International Air Transport Association, 2017). As a consequence of these accidents, 268 people lost their lives in commercial aviation accidents in 2016. Consulting analyses to determine the causes or contributing factors of today's accidents, one will find Shappell and Wiegmann's (1997) oft-cited number stating that about 80% of the accidents are due to, or related to, human error. Although these authors used this statistic very carefully, the 80% estimate is nowadays often used as a catchy and generalized statistic, which is insinuating a one-dimensional and mono-causal relation between accidents and human error (Hoermann & Lorenz, 2009); see also Bengler, Winner, and Wachenfeld (2017) for a critical consideration in the automotive domain. Modern accident analyses see so-called human errors as a need to dig deeper into the systemic and organizational framework (Hoermann & Lorenz, 2009), as they are considered the results of poor system design (Norman, 1983), while older approaches saw human errors as foremost attributable to pilot performance.

1.1 Crew resource management was born

The important role that humans play in the sociotechnical system of an aircraft cockpit as part of a team was first verbalized after NASA held a workshop on "Resource Management on the Flightdeck" (G. E. Cooper, White, & Lauber, 1980; Helmreich, Merritt, & Wilhelm, 1999), which was widely attended by airline representatives and researchers. As a result of this discourse, researchers and airlines acknowledged that pilots' nontechnical skills, such as communication (Kozlowski, Gully, Nason, & Smith, 1999; Salas, Cooke, & Rosen, 2008; Straus & Cooper, 1989), workload management (Smith, 1979), and decision-making (Orasanu, 1993; Orasanu, Dismukes, & Fischer, 1993), are contributing factors or causes of accidents more often than technical malfunctions (e.g., Gontar, Fischer, & Bengler, 2017; Gontar & Hoermann, 2014, 2015b; Gontar, Hoermann, Deischl, & Haslbeck, 2014). This surprising fact led to a rethinking process of the whole aviation industry. As a result, efforts were undertaken to focus more on the human than on the technical advances; this approach was first titled *cockpit* resource management before changing to *crew* resource management (CRM) (Helmreich & Foushee, 2010). This marks the point in time where a new training and research era was born that would greatly reduce the occurrences of incidents and accidents. In the subsequent years, several airlines and also military agencies followed the example of United Airlines, who was the first to implement CRM training, and introduced respective courses as part of their initial and recurrent training (Carroll & Taggart, 1987).

1.2 Crew resource management was regulated

CRM as part of initial and recurrent pilot training was later defined in the International Civil Aviation Organization's (ICAO) Annex 6 Part 1 Chapter 9.3.1 (International Civil Aviation Organization, 2001) and brought into law under Title 14 of the Code of Federal Regulations (14 CFR) part 121 in the United States (Federal Aviation Administration, 1996) and under Commission Regulation (EC) No 859/2008 OPS 1.940 in the European Union (European Commission, 2008) as operating rules, which are currently regulated under European Aviation Safety Agency (2015b); see also Farrow (2010) for an in-depth description on the regulatory evolution. Operating rules, in contrast to airworthiness standards, are continuously applicable (Abbott, 2010). This means, whenever the regulation changes, the aircraft operator has to make sure that the updated rules are applied (e.g., European Aviation Safety Agency, 2015b).

Concerning CRM, the Federal Aviation Authority (FAA) directed an amendment of the Federal Aviation Regulations (FARs) to make CRM training obligatory (Birnbach & Longridge, 1993) in the late 1980s. At this point, the FAA was already assured "…that in today's technological and operational environments, it is not sufficient to provide training that focuses only on an individual pilot's technical competence" (Birnbach & Longridge, 1993, p. 270). As a consequence, the FAA regulated CRM training in their rulemaking, 14 CFR Part 121 Section 121.404 (Federal Aviation Administration, 1996), such that:

> "After March 19, 1998, no certificate holder may use a person as a flight crewmember, and after March 19, 1999, no certificate holder may use a person as a flight attendant or aircraft dispatcher unless that person has completed approved crew resource management (CRM) or dispatcher resource management (DRM) initial training, as applicable, with that certificate holder or with another certificate holder."

To help operators implement such new regulations, the FAA as well as the European Aviation Safety Agency (EASA) publish a so-called advisory circular (AC) and acceptable means of compliance (AMC). In those documents, the regulators give further details on how specific regulations have to be utilized and implemented into flight training and operation (Abbott, 2010).

Regarding CRM training, the FAA issued and constantly updated its ACs. In the current form (U.S. Department of Transportation, 2004), the AC asks operators to implement training with three major components: (1) Initial Awareness Training, (2) Recurrent Practice and Feedback, and (3) Continuing Reinforcement. It is suggested that component (1) is implemented in classroom presentations, lectures, group exercises, and role-play. Further, surveys among trainees bring the benefit of identifying operational problems and personal attitudes (see chapter on *trainee characteristics*) (U.S. Department of Transportation, 2004). Component (2) is based on recurrent training, with CRM skills manifested during briefings, debriefings, and feedback. Preferably, feedback is given to the full crew and is based on video recordings of the trainees' simulator mission to identify specific behaviors that show the strengths and weaknesses of the crew. The use of video recordings further allows the trainees to see themselves from a third-person view and assess how they might act upon their colleagues (see chapter on *training design*) (U.S. Department of Transportation, 2004). Component (3) emphasizes the fact that training is ineffective if it is not constantly lived during real operation (see chapter on *work environment*) (U.S. Department of Transportation, 2004). It is important that an airline facilitates an atmosphere that fosters the very principles of CRM training, such as common decision-making, speaking-up, or simply said: *working as a team*.

1.3 Positive transfer of training is the key

As one of the oldest questions addressed in industrial and organizational psychology, positive transfer of training (Bell, Tannenbaum, Ford, Noe, & Kraiger, 2017) is the crucial element when organizations develop training for their employees. Studied for about 100 years now, training transfer is defined as "the degree to which trainees effectively apply the knowledge, skills, and attitudes gained in a training context to the job" (Baldwin & Ford, 1988, p. 63). Extensive efforts were undertaken to achieve positive transfer in all different domains. The oft-cited value that only 10% of the training actually transfers to on-the-job behavior might serve as a catchy introduction, but was just a speculation (Georgenson, 1982) and not based on any empirical data (Saks, 2002). However, as found in a survey by Saks (2002), only about 50% of training invests result in positive training transfer. Although this estimate is not as worrisome as the 10% estimate, lack of positive training transfer might still result in unnecessary expenditure. In addition, high-reliability organizations may also suffer from increased risk when training does not sufficiently prepare pilots for real-world situations.

To allow for a structured approach on improving training transfer, Baldwin and Ford (1988) performed an extensive literature review and developed a model that shows the influences on the transfer process clustered into three major categories (see Figure 1.1).

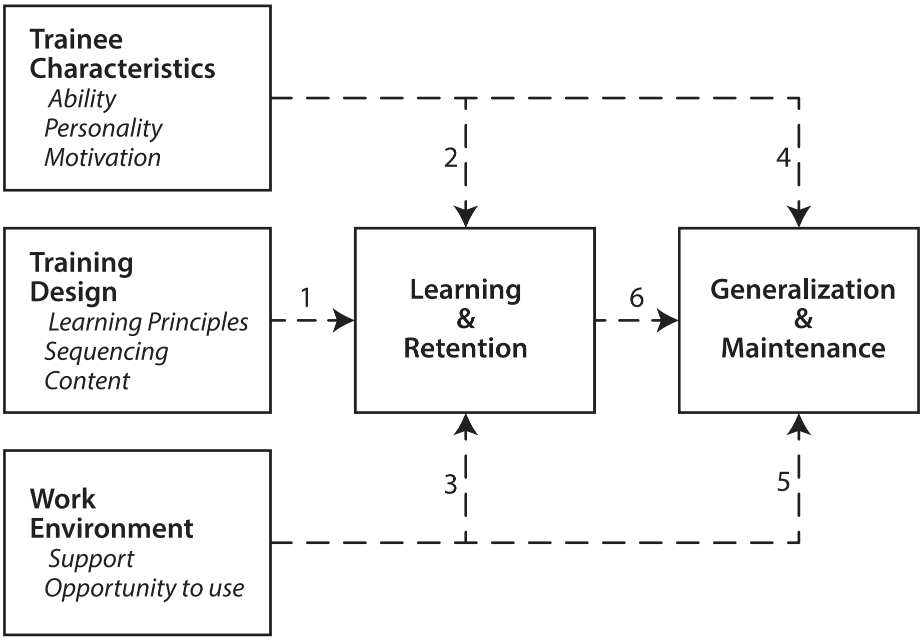

Figure 1.1: Training transfer model according to Baldwin and Ford (1988)

This model was chosen because most transfer research falls into these three categories (Alvarez, Salas, & Garofano, 2004; Burke & Hutchins, 2007) and extensive and successful efforts were undertaken to validate it over the last 30 years (Ford, Baldwin, & Prasad, 2017). In the following, the training transfer model will be described in greater detail, however, it is explicitly pointed out that this thesis does not aim to validate this model but rather to use the model as a framework to examine potential negative influences on the training transfer in pilot training.

Baldwin and Ford's (1988) transfer model defines three training input variables found to be of high relevance when it comes to training transfer, as they are hypothesized to directly influence learning and retention of training contents (linkage 1–3) and, therefore directly and indirectly, generalization and maintenance.

Trainee characteristics

Placed on the upper left of Figure 1.1, trainee characteristics refer to latent variables that are connected with the person attending the training and his/her abilities, personality, motivation, (Baldwin & Ford, 1988; Blume, Ford, Baldwin, & Huang, 2010; Facteau, Dobbins, Russell, Ladd, & Kudisch, 1995; Weissbein, Huang, Ford, & Schmidt, 2011), self-efficacy (Chiaburu & Marinova, 2005; Gegenfurtner, Veermans, & Vauras, 2013), perceived utility (Nikandrou, Brinia, & Bereri, 2009), and were found to influence learning itself along with generalization and maintenance (Burke & Hutchins, 2007). Generally speaking, within pilot selection processes (Carretta & Ree, 1994), attempts are made to control several aspects attributed to trainee characteristics. Pilot selection defines the process of comparing the trainees' characteristics to a predefined set of job requirements (Goeters, Maschke, & Klamm, 1998) by using reliable and valid selection tests (Damos, 1996). Although selection tests are validated via training success (Martinussen, 1996), some of the trainees' characteristics may change over time and are thus not exactly predictable in the selection process (Helmreich, Sawin, & Carsrud, 1986). One can imagine that motivation (intrinsic and extrinsic) might change with a pilot's career path, with the daily routine, or simply because he/she was expecting and imagining a different job profile—to name just a few examples.

Work environment

Placed on the lower left of Figure 1.1, Baldwin and Ford (1988) highlighted the work environment as a predictor of training transfer. They initially found evidence of *support systems* and the *opportunity to use the learned* being major predictors of successful training transfer (Schindler & Burkholder, 2016). Support systems include colleagues or supervisors being available for discussion and questions about the training context (Burke & Hutchins, 2007). Further, the opportunity to use newly learned skills is seen as one of the most important factors influencing training transfer (Ford, Quinones, Sego, & Sorra, 1992; Schindler & Burkholder, 2016).

In aviation, opportunities to use learned skills are often restricted to the training context; especially when it comes to handling technical or procedural problems and malfunctions, which must not be simulated during real operation according to ICAO Annex 6 Part 1 Chapter 4.2.4 (International Civil Aviation Organization, 2001). While emergency situations rarely arise during normal operation, exactly these problems can cause incidents or accidents, requiring appropriate problem-solving strategies from the pilots. The usage of CRM skills when handling problems, however, is only trainable in simulator sessions. This fact puts particular importance on the design of the training, which directly influences the training outcome and, thus, training transfer.

Additional research has evaluated further influences on training transfer arising from the work environment (Burke & Hutchins, 2007). Transfer climate, which can be seen as the degree to which an organization and its management encourage or discourage trainees from making use of their newly trained skills, was found to directly impact training transfer (Brinkerhoff & Montesino, 1995; Richman-Hirsch, 2001; Schindler & Burkholder, 2016); see also the literature on organizational culture (e.g., Lim & Nowell, 2014). The latter can be seen as linked to the *opportunity to use*, but from an organizational point of view. Speaking from an operational point of view, accountability—the degree to which an organization holds trainees responsible for using newly trained skills—was also found to be positively related to training transfer (Burke & Hutchins, 2007).

Training design

Because pilot training for abnormal situations is normally restricted to simulators, great importance is attached to its design (placed in the middle left of Figure 1.1). That is, it constitutes the main topic of this cumulative thesis. Besides training content (Baldwin & Ford, 1988), which is picked up by Burke and Hutchins (2007) and titled content relevance, the latter also point to the importance of performance feedback to achieve positive training outcomes and, thus, training transfer (see also Karl, O'Leary-Kelly, & Martocchio, 1993). Burke and Hutchins (2007, p. 274) conclude that trainees "…must see a close relationship between training content and work tasks to transfer skills to the work setting", while Baldwin and Ford (1988, p. 67) emphasize that "…feedback is a critical element in achieving learning and that timing and specificity are critical…" How these two constructs relate to pilot training in the civil and commercial aviation domain is explained in the following two subchapters.

1.4 Training has to feature reliable performance feedback

Following the rules set down by the FAA and EASA, CRM training modules were designed to include aspects of communication, coordination, workload management, situation awareness, and decision-making (European Aviation Safety Agency, 2015a). To assess a potentially positive effect of the newly introduced CRM training, trainees' performance had to be assessed (for details see Salas, Burke, Bowers, & Wilson, 2001). Further, reliable performance assessment was needed to give appropriate feedback to the trainees and also to form a basis for training modifications in terms of needs assessments (Goldsmith & Johnson, 2002)—all important factors to facilitate positive transfer of training. In addition, especially high-risk organizations depend on reliable performance assessments of their employees to ensure a sufficient level of safety.

To measure nontechnical performance, behavioral marker systems were developed (Helmreich, Wilhelm, Kello, Taggart, & Butler, 1990). That is, performance assessment is based on directly observable behaviors, such as "the pilots discussed the upcoming course of action" rather than "proactive task management," to reduce the variance from different mental models on the side of the observer. Even though behavioral marker systems were designed to serve as objective performance criteria and to improve the standardization among judges, they all require human observers, who inevitably add variance to the assessment process. This variance on the side of the human raters in the nontechnical performance assessment can reduce the reliability of the performance feedback given to the trainees. Feedback reliability refers to interrater reliability on the side of the trainers, describing the extent to which the assessment of nontechnical skills in pilot training is consistent across judges who rate the behavior or the performance of pilots. High interrater reliability is given when a certain level of performance is rated with roughly the same grade from every observer or trainer (Shrout & Fleiss, 1979).

Feedback to the trainee can become worthless if one trainer enforces a certain behavior while another trainer discourages it (low interrater reliability). One can easily imagine that unreliable performance feedback can be very detrimental to the skill development of each trainee as he/she cannot attribute the performance assessment to his/her own behavior. Possible consequences might be that the motivation of the trainee will be lowered or that the trainee is assessed as "ready to fly" when he/she is not (Gontar & Hoermann, 2015a; Holt, Hansberger, & Boehm-Davis, 2002). While the latter introduces a safety hazard, reduced motivation as part of the trainee characteristics (see Burke & Hutchins, 2007; Chiaburu & Marinova, 2005; Facteau et al., 1995) could greatly decrease the transfer of training. Interrater reliability as a basis for a reliable feedback is consequently an aspect that has to be treated very carefully and might even be seen as a precondition for training validity.

1.5 Training has to feature content relevance and validity

Even if trainers can assure reliable performance assessment in training, it does not necessarily mean that the pilots can transfer their learned skills to real operations or real malfunctions. Whether positive transfer of training can be expected also depends on the content relevance and validity of the training (Axtell, Maitlis, & Yearta, 1997; Burke & Hutchins, 2007; Holton, Bates, & Ruona, 2000). An example could be upset prevention and recovery training. In reality, such situations would load pilots with high forces, as the aircraft might maneuver into unusual bank and pitch angles, resulting in increased roll or pitch rates. However, such training is currently provided in full flight simulators that are—per design—not able to expose pilots to such high loads required to reflect reality. If pilots experience a real upset situation during flight, they may be unable to react adequately, as they would have experienced different aircraft behavior and vestibular feedback from training: an effect often labelled *negative training* (Foster et al., 2005).

Although such upset prevention and recovery training could feature a high degree of interrater reliability in the performance assessment, the validity of the training would not be given. The training in this example is built upon assumptions about flight behavior that are not true. Although this example seems rather radical, pilots will experience such situations, in small nuances, several times during their career—maybe not related to the aircraft's flying behavior, but to the airline operation. Pilots might always have sufficient time to solve a problem in the simulator, but in reality, there might be environmental constraints that force them to develop new strategies that they have not previously applied. Only when pilots are provided with tools and training that are applicable in reality, thus featuring high content validity, their behavior will change (Helmreich & Foushee, 1993) and positive training transfer can be expected (Burke & Hutchins, 2007).

2 A peek on previous research

A large body of research is available on the transfer of training. However, in relation to pilot training, some issues remain unresolved. The following two subchapters give a short introduction to the respective literature. Nevertheless, it is noted that this can only give a very short introduction to the respective research areas and that the extensive and detailed literature review is given in the attached papers. The goal of this research, however, is not to analyze the impact of the mentioned constructs on the degree of training transfer, but to analyze whether there are indications that feedback reliability and content relevance are not sufficiently met in current training.

2.1 Reliability of CRM skill assessment is still an open issue

With the development of behavioral marker systems several decades ago (Butler, 1991; Flin & Martin, 2001; Helmreich et al., 1990; Seamster, Hamman, & Edens, 1995), and the explicit advice to ensure sufficient interrater reliability by using observable markers rather than generic items (Flin & Martin, 2001), one would expect high reliability in everyday training feedback. However, a considerably large amount of research was conducted in several domains showing that high interrater reliability can never be taken for granted (e.g., Arora et al., 2011; Kontogiannis & Malakis, 2013; Sevdalis et al., 2008; Yule et al., 2009). Research shows an influence of raters themselves (expertise, motivation, personal interpretation) (Yule et al., 2008, 2009), contextual factors (who was rated in which scenario during which task) (Mitchell et al., 2012; O'Connor, Hoermann, Flin, Lodge, & Goeters, 2002), rating dimensions (social or cognitive skills) (Dedy et al., 2015; Mishra, Catchpole, & McCulloch, 2009), and the scales (five-point or pass/fail) used (O'Connor et al., 2002) on the reliability of the nontechnical skill rating. Although absolute reliability can hardly be reached when involving human raters (Weber, Mavin, Roth, Henriqson, & Dekker, 2014), it is very important to be aware that reliability might not even be as good as one would expect. Knowing the limits of reliability is very important to reduce the potential of inadvertent overreliance in CRM assessments.

Another important aspect of a reliable performance assessment is the pilots' self-assessment of their performance. When we asked pilots in one of our studies to assess their own and their peers' skills and then compared their assessments to the instructors' ratings, we found considerable disagreement (Gontar & Hoermann, 2014). One could argue that it is not the pilots' self-assessment or peer assessment that counts in order to pass the training, but rather that of the instructor. However, all pilots receive feedback using the same behavioral marker system and would benefit if their own understanding of the rating dimensions was the same as the instructors' (Gontar & Hoermann, 2014) to allow positive training outcomes—and, thus, transfer.

Among all the skills taught and assessed in CRM training, communication occupies a special position. Communication is the means by which pilots facilitate team building (Kanki & Palmer, 1993), coordinated teamwork (DeChurch & Mesmer-Magnus, 2010), shared situation awareness (Orasanu, 1994; Orasanu, Fischer, & Davison, 1997), leadership (Palmer, Lack, & Lynch, 1995)—essentially all aspects of CRM (Kanki & Smith, 2001). Communication allows the unlocking of "the cognitive power of a crew" Orasanu (1994, p. 277), as it is "the glue that binds participants together in group interaction or team tasks. It is a transparent medium through which group work is organized and accomplished" (Orasanu et al., 1997, p. 2). In high-risk environments, communication has to be very effective and efficient, especially in abnormal situations (Orasanu et al., 1997), fulfilling three major needs (Orasanu, 1994): sharing information, directing actions, and reflecting thoughts. As communication is shown to be a very important factor in CRM, and essentially the medium through which CRM within a crew can be facilitated, it is not surprising that it affects a variety of team performance aspects (Gontar, Fischer, & Bengler, 2017; Salas et al., 2008; Salas, Sims, & Burke, 2005). Hence, to enhance safety, it is of utmost importance to reliably assess and provide feedback on the communication skills of pilots (Gontar, Fischer, & Bengler, 2017).

Several studies and approaches have been considered, but they have achieved only mediocre success in finding a gold standard by which to evaluate pilot communication (e.g., Bourgeon, Valot, & Navarro, 2013; Krifka, Martens, & Schwarz, 2004; Mosier & Fischer, 2010). Some of those studies used an approach grounded in speech act theory (Austin, 1962; Searle, 1969). This approach is very time consuming, not real-time capable, and not applicable in pilot training (Cooke & Gorman, 2009) because human coders have to assign speech acts from a large inventory to every single utterance. This asked for the implementation of content-free analyses (e.g., Cooke & Gorman, 2009; Fischer, McDonnell, & Orasanu, 2007), to describe pilot crew communication.

Gorman, Cooke, Amazeen, and Fouse (2012) recently used a non-linear method called cross-recurrence quantification analyses (Marwan, Carmen Romano, Thiel, & Kurths, 2007; Marwan & Kurths, 2002; Webber & Marwan, 2015) to measure the communication behavior from three-person teams showing promising results. The foremost advantage of this method is that it does not assume any linear interaction but also covers non-linear relations between systems. We previously used this method to visualize and quantify pilots' coordination strategies using eye-tracking data (Gontar & Mulligan, 2016) and urged the community to further implement non-linear approaches in the analysis of human interaction and coordination behavior as the assumptions for linearity might not always be given.

Although the research community offers several different approaches to reliably assess and provide feedback on the communication skills of pilots, there is no sufficiently precise and, at the same time, practicable approach to be included in airline training that is capable of distinguishing between different levels of pilot crew performance. Properly implemented, cross-recurrence quantification analyses could bridge this gap.

2.2 CRM skills can only transfer when the training is content valid

As discussed at the beginning of this work, reliability in the assessment is necessary, but not yet sufficient, to ensure positive transfer of training. Another necessary aspect is the content relevance and validity. The basis for pilot training is the operation manual developed or adapted by the airline. While the aircraft manufacturer publishes the general aircraft operation manual, which defines how the airplane is best operated, each airline might want to modify several parts of it to fit their particular operational needs (Degani & Wiener, 1991). The degree of this modification is influenced by different factors; thinking of the different requirements for long-haul and short-haul flights or for passenger and cargo flights, it seems reasonable that there cannot be a *one-size-fits-all* solution.

When pilots undergo their initial training today, the training context is clustered according to technical togetherness rather than operational togetherness (Barshi, 2015). Pilots might start by studying the details of the braking system, with its related hydraulic systems, and afterward focus on the electric system, before studying the auto-flight system. In his work, Barshi (2015) suggests another training approach of structuring pilot training according to flight phases rather than technical systems. The training—as he calls it, "comprehensive line-oriented flight training"—is designed to follow the training principles of Healy, Kole, and Bourne (2014) to better train pilots from the very beginning of their career for real operation instead of an ideal environment. Analyzing whether the ideal is met in real operations, as is assumed by training and operating procedures, an outstanding piece of work was accomplished by Loukopoulos, Dismukes, and Barshi (2009). These authors observed hundreds of flights during real operation and derived a very explicit structure and overview on the assumptions made in the ideal world of training and procedures, pointing out the differences to the real operation (see also Loukopoulos, Dismukes, & Barshi, 2003). Operation manuals and training are based on the following assumptions: (1) linearity, describing the consecutive order of tasks that have to be accomplished by the pilots one by one, (2) predictability, describing the possibility of pilots anticipating tasks in terms of their occurrence and their content, and (3) controllability, describing the possibility of pilots controlling the execution of the task independently of anything else (Loukopoulos et al., 2009).

The problem here is that during training pilots will not experience extreme time pressure and will thus be able to work through the procedures in a structured and linear way. Given enough time, pilots will have the feeling that they can control the execution of the tasks according to their needs and capabilities. In terms of predictability, the problem is even worse. All pilots have to be trained and checked in the same way within one airline to achieve the same knowledge basis, being interchangeable, and also to maintain the same level of proficiency and thus safety. Hence, the pilots will know exactly what will happen during their four-hour training and check shift. It is desirable, of course, that pilots show up well prepared to make the most of the training time, but this conflict leads to a situation in which pilots do not experience real unforeseen events during training (Casner, Geven, & Williams, 2013); even worse, pilots might even believe that they can handle unforeseen events because they succeed in the training courses. In fact, in simulator training, pilots mentally simulate the whole expected scenario and think through several options beforehand. In reality, this predictability is, of course, not a given, as technical and operational problems often occur without warning. In reality, pilots have to handle unforeseen events, but training seems not to feature this (Casner et al., 2013).

Even though one can imagine that in a complex system such as aviation, the mentioned assumptions of linearity, predictability, and controllability do not always hold true, it is also understandable that assumptions have to be made. For example, the assumption of linearity has to be made, because this is the nature of written language, with one sentence following another, and written procedures might not be visualized differently (Linell, 1982). However, Loukopoulos et al. (2009, p. 43) conclude "…that the design and training of procedures are not sufficiently robust to deal with the complexities of real-world operations…" The problem that results from everyday challenges not being covered by appropriate procedures and training is that pilots have to find their own solutions (Loukopoulos, Dismukes, & Barshi, 2001) but are not trained to handle such situations. Engaging in developing their own solutions and strategies might create a large variability in pilots' response behavior, require much effort from them, and increase their workload (Orasanu et al., 1993)—all aspects that can ultimately lead to reduced safety.

In this context, interruptions, which manifest a violation of all three major assumptions upon which flight operation manuals and training are based, were found to have a large negative effect on the overall level of flight safety (Loukopoulos et al., 2001). Although this fact has been known for years, it seems that not enough effort has been undertaken to help pilots deal with interruptions. Interruptions and their potential effects are not an explicit part of pilot training to raise the awareness of potential effects interruptions might have, nor are there any procedural changes in place to minimize the number of interruptions or their consequences. As interruptions are known to increase workload (Gupta, Li, & Sharda, 2013; Weigl, Antoniadis, Chiapponi, Bruns, & Sevdalis, 2015; Weigl, Müller, Vincent, Angerer, & Sevdalis, 2012; Zijlstra, Roe, Leonora, & Krediet, 1999), affect the emotional state of the operator (Bailey & Konstan, 2006), and increase the error probability of the operator (Elfering, Grebner, & Ebener, 2015; Flynn et al., 1999), it seems extraordinarily important to highlight these effects in the aviation domain. As long as training does not take the real into account, it suffers in terms of content relevance and, hence, validity. Consequently, the ostensible high performance of the crew shown in simulator or training missions might not transfer to the real operation, resulting in an overreliance on the safety of the aviation system.

3 Method

3.1 Research questions

The literature review (Gontar & Hoermann, 2015a) revealed several aspects of the current research that are not sufficiently considered when it comes to the interrater reliability evaluation of CRM training in terms of nontechnical skill ratings made by experts. The following research questions are formulated:

1. How reliably can expert trainers assess pilots' nontechnical performance using assessment tools with which they are already familiar?
2. Which dimensions of pilots' nontechnical skills feature high or low interrater reliability?
3. What factors (e.g., flight scenario, rating scale) influence interrater reliability?

The goal of the first paper dedicated to interrater reliability was to estimate the reliability of very experienced raters when using a familiar rating tool to identify dimensions of very poor reliability and to give a systematic overview on factors that influence the rating reliability.

As communication is considered the most important aspect of CRM, and because it has also been shown to have particularly low interrater reliability (Gontar & Hoermann, 2015a), the aim was to develop an alternative to the classical behavioral marker approach. Looking into communication assessment in training, one can identify further research needs that are addressed in the paper of Gontar, Fischer, and Bengler (2017) in terms of attempting to answer the following question:

4. Is there a reliable, objective, and transparent method to evaluate crew communication during training that is capable of distinguishing between different levels of crew performance?

The goal of this paper was to suggest a more reliable approach for measuring crew communication than expert ratings. To achieve this goal, we investigated the feasibility of implementing and modifying a nonlinear method—namely, cross-recurrence quantification analysis—to be capable of distinguishing between different levels of crew performance. We used a nonlinear approach, as the interaction between crew members was already shown to follow nonlinear behavior when we adapted this method to measure the degree of joint visual attention in pilot teams (Gontar & Mulligan, 2016; Mulligan & Gontar, 2016).

Regarding training validity, Gontar, Schneider, et al. (2017) focused on whether the assumptions made by procedures and processes that are the basis for training transfer actually reflect reality.

5. Does the real operation, specifically the turnaround, follow the assumptions made in procedures and training?
6. How many interruptions occur during the turnaround, and how do they influence pilots' workload and their committing of errors?
7. Which other factors are predictive of pilots' workload during the turnaround?

The goal of this paper was to analyze whether interruptions occur during turnaround processes, and how they influence pilots' perceived workload and their committing of errors. We conclude with recommendations to reduce the number of interruptions to limit the violation of the assumptions made by the procedures and training.

As pilots are mentally prepared for their upcoming training mission during normal airline training, it seems especially important to analyze their behavior in scenarios that are unexpected and that they have never seen or mentally walked through. This will provide an understanding of their reaction in real-world situations. The increase of cyberattacks on various systems (de Zan, d'Amore, & Di Camillo, 2016; Elias, 2015; Wilshusen, 2013) was found to be an appropriate context, as it is currently not even considered in pilot training. The research needs addressed by Gontar et al. (2018) are summarized as follows:

8. How does a cyberattack influence pilots' performance, their visual information acquisition, their perceived workload, and their trust in the system?
9. How do alarms moderate these effects?

The goal of this research was to determine how well pilots could handle a situation for which they have not been trained. Another aspect was to obtain first knowledge about pilots' reactions to cyberattacks. With this knowledge provided, procedures and training can be amended before an incident or accident forces the airline industry to concede that pilots are not trained for these new challenges. Early research on this topic allows us to proactively—which is not very common in aviation—tackle new hazards before they even harm us.

3.2 Research conducted and its content

To answer these research questions, four experiments were conducted involving a total of 339 pilots (see Table 3.1). Given the challenges of external validity, we only used licensed pilots as participants, as using students or novices as participants could have jeopardized the validity of the results (see Druckman & Kam, 2009; Oakes, 1972; Sears, 1986). Furthermore, to avoid self-selection bias, we recruited our participants (Experiments 1–3) on a random basis and did not invite volunteers (Rosenthal & Rosnow, 1975). Only in Experiment 4 were we unable to recruit on a random basis and hence used volunteers. However, as the scenario was designed to be completely novel and unforeseeable for the pilots, we did not anticipate any problems with a potential self-selection bias.

Table 3.1: Overview of experiments conducted and corresponding papers included

Paper	Goal	Participants	Environment
#1	…to assess the interrater reliability of the best raters under ideal conditions	37 type-rating examiners (random)	Classroom & video recordings
#2	…to assess the communication behavior of pilots with a new method of assessment	120 type-rated pilots (random)	Level-D full motion Airbus simulator
#3	…to assess the extent to which real operation differs from the ideal in training	160 type-rated pilots (random)	Real operation at airport
#4	…to assess the extent to which an unforeseen event influences pilot behavior	22 pilots (volunteers)	Static & generic simulator

Note. Each embedded paper corresponds to one experiment.

Experiment 1 was conducted in a classroom because we wanted to test 37 instructor pilots at the same time under the best of possible observation conditions. That is, the instructor pilots observed crew behavior based on video recordings as opposed to actual training situations where they are sitting in a rather dark simulator with several other duties in parallel. Looking at Experiment 2, we were able to maintain the highest possible external validity level. Certified full flight simulators are *copies* of aircraft cockpits and are permanently used in pilot training; they also feature the only environment to evoke technical malfunctions. In this context, it is pointed out that within the scope of Experiment 2, two separate scenarios were presented: one that focused on technical malfunctions and handling abnormal procedures to which this thesis refers, and a second that focused on manual flying skills. The latter is not addressed in this thesis, but details can be found in Haslbeck (2017). Experiment 3 was executed during real operation at a major European airport, and thus it features the highest possible external validity. For the last experiment, we chose a generic simulator that would allow us to accept pilots with different aircraft type ratings (e.g., Boeing, Airbus). In total, we were able to collect data from 339 individuals, allowing for a considerable high degree of generalizability.

4 Results

Each paper of this publication-based thesis presents the results of one of the experiments. Paper #1 focuses on a critical issue for positive training outcomes and, consequently, transfer: the reliable performance assessment as the basis for feedback. It offers a rigorous literature review on different methods of measuring interrater reliability, reports the results from the interrater reliability study, and provides an overview of the factors that have the greatest influence on interrater reliability. It concludes that the research community should foster positive transfer of training and it provides recommendations to the industry for the assessment of pilots' nontechnical skills. Communication was found to be especially affected by low interrater reliability.

While it is one of the most important skills to be learned in CRM, Paper #2 proposes a novel nonlinear approach—cross-recurrence quantification analysis—to analyze pilots' communication behavior. We compared it with commonly used methods grounded in speech act theory and conclude that this new method can be used to reliably assess communication behavior. It might be implemented in pilot training as an alternative to the currently used expert ratings.

Paper #3 focuses on the content relevance of training and thereby on the assumptions underlying current airline procedures and training. It examines whether the assumptions made reflect the current practice within the airline operation. We conclude that current training does not reflect reality, as most assumptions are violated in reality. We recommend two approaches for the airline industry to cope with this specific problem.

Looking at content relevance, we report in Paper #4 how we put pilots in a completely unforeseeable situation where they were faced with a cyberattack—a scenario that none of the pilots has ever trained for or cognitively simulated before. We analyzed how pilots reacted in this unforeseen situation and how it impacted their behavioral and psychological state. We conclude that future training should feature unforeseen events, and we recommend incorporating such training as an immediate action to foster training transfer for unforeseen events.

4.1 Interrater reliability in the assessment of pilots' CRM skills (Gontar & Hoermann, 2015a)

Gontar, P., & Hoermann, H.-J. (2015). Interrater Reliability at the Top End: Measures of Pilots' Nontechnical Performance. *The International Journal of Aviation Psychology*, *25*(3-4), 171–190.

> »*If you can not measure it, you can not improve it.*«
> —*Lord Kelvin*

As pointed out by Lord Kelvin, it is crucial to reliably measure whatever constructs are relevant in a specific context in order to improve them. Regarding pilots' nontechnical skills, it is not only about improving them but also providing appropriate feedback to the trainee to ensure sufficient training transfer. After CRM training was introduced in the early 1980s, several behavioral marker systems have been developed and are currently used by various airlines. However, it seems that interrater reliability is not always a given.

In this experiment, a total of 37 type-rating examiners from a major European airline examined video recordings to rate the performance of four flight crews showing different levels of performance. The raters used three different behavioral marker sets: one based on the Line Operations Safety Audit (LOSA) approach and landing sheet, and the other two based on modified nontechnical skills (NOTECHS) tools. To quantify the degree of within-group agreement and interrater reliability, we calculated r_{wg} and two-way mixed-model intraclass correlation coefficients $ICC(3)$.

Results showed acceptable within-group agreement for the crew showing the best performance in this sample. Ratings for the medium-performing crews showed unacceptable agreement when rated with the two modified NOTECHS tools. Interrater reliability across all rating dimensions and rating tools was below the necessary standard. Comparing different scales, we found that pass/fail scales, compared to four- and five-point scales, showed lower agreement. Looking at the different rating dimensions, we found that social dimensions (communication, leadership, and teamwork) were rated with less reliability than cognitive aspects (work organization, situation awareness, and decision-making). Further, we found systematic differences in the rating for captains and first officers, as well as between different crew performance levels.

The analyses of the most experienced raters within an airline showed insufficient agreement and reliability in the rating of pilots' nontechnical skills. We showed that the consistency of a nontechnical skill rating is highly dependent on the target being rated—namely the performance and the context—the rating dimension (e.g., communication, teamwork), the scale used, and the rating tool.

4.2 A nonlinear approach to measuring crew communication (Gontar, Fischer, & Bengler, 2017)

Gontar, P., Fischer, U., & Bengler, K. (2017). Methods to Evaluate Crew Communication in a Training Environment: Speech Act Based Analyses vs. Cross Recurrence Analysis. *Journal of Cognitive Engineering and Decision-making*, *11*(4), 337–352.

»*Two monologues do not make a dialogue.*«
—*Jeff Daly*

Flying an airplane means controlling it. Nowadays, controlling does not merely refer to the control surfaces but rather to managing the auto-flight system, monitoring all aircraft systems, coordinating the tasks within the pilot crew, and making sound decisions in case of any malfunctions or operational constraints. Communication facilitates everything in the cockpit. However, as communication was found to be rated with considerable disagreement among the most experienced raters, a new method had to be developed to reliably assess the communication behavior of pilots during training.

We ran an experiment with 120 licensed and randomly selected pilots, presenting them with two malfunctions on their final approach. Besides other research questions, this piece of work used the communication data of the pilot crews to implement a new method of assessing communication; we implemented a nonlinear cross-recurrence quantification analysis—a method previously shown to be able to measure the degree of coordination in pilot teams (Gontar & Mulligan, 2016; Mulligan & Gontar, 2016). This method is compared to a commonly used method based on speech act theory.

Using the well-established speech act theory-based approach, we found that the best crews have a significantly higher anticipation ratio than their poorly performing colleagues. This means that the best crews are very likely to share information without a team member having to request it. Applying the developed cross-recurrence quantification analysis, we found that the best crews show a higher adaptation rate in their communication behavior than the poorly performing crews. Further, the poorly performing crews are characterized by significantly higher instances of interrupting each other.

The analyses showed promising results using the speech act theory-based approach. However, the coding of the single speech acts is very time consuming and hence not applicable for airline training. As a method that does not rely on human coding, cross-recurrence quantification analysis was also able to identify communication patterns linked with high and low crew performance. Given the poor reliability of nontechnical skill ratings in the dimension of communication, cross-recurrence quantification analysis constitutes an alternative that could be used in pilot training, possibly facilitating objective and transparent feedback.

4.3 Differences between ideal and real (Gontar, Schneider, et al., 2017)

Gontar, P., Schneider, S. A. E., Schmidt-Moll, C., Bollin, C., & Bengler, K. (2017). Hate to Interrupt You, but… Analyzing Turn-Arounds from a Cockpit Perspective. *Cognition, Technology & Work*, *19*(4), 837–853.

> »*Euclid taught me that without assumptions there is no proof.*
> *Therefore, in any argument, examine the assumptions.*«
> —*Eric Temple Bell*

Aircraft manufacturers and operators make three major assumptions when it comes to the design and implementation of training and procedures: linearity, predictability, and controllability. Given these fundamental assumptions in an idealized world, the question is whether the assumptions are met in reality. If this is not the case, the content validity of some aspects of pilots' CRM training is questionable. Interruptions, which are an inevitable violation of the mentioned assumptions, are not sufficiently investigated in aviation, especially not in turnarounds, where we did not find a single experimental study.

To assess the validity of the assumptions made by procedures and training, we conducted an experiment with 80 randomly selected, fully certified two-pilot crews (160 pilots in total). In this observational study, we counted and classified interruptions experienced by the pilots, measured pilots' subjectively perceived workload, and counted and classified their committed errors. To quantify potential influences on pilots' perceived workload that might arise from external factors such as weather, flight delay, and so on, we collected these data and implemented them in a multiple linear regression model.

Results showed that pilots had to handle about eight interruptions per turnaround. The overall perceived workload does reflect a comparable level as we found it for manual flying tasks. Interruptions originating from external sources were found to be predictive of pilots' perceived workload. Taking into account the external factors, we found that poor weather conditions had an even higher impact than interruptions on the workload perceived. In total, we observed 64 errors committed by pilots, but all were successfully recovered during our observation period. We did not find any relationship between interruptions and errors.

This study demonstrates the weakness of the assumptions made in procedures and training. Interruptions, which were shown to occur frequently, disrupt pilots during one task and often force them to squeeze in another task simultaneously. When interruptions occur, pilots no longer control their workflow, and they cannot predict when they will face another disruption of their task. We conclude that linearity, predictability, and controllability are not met during turnarounds, and we suggest that training should take this into account.

4.4 Facing the unforeseeable (Gontar et al., 2018)

Gontar, P., Homans, H., Rostalski, M., Behrend, J., Dehais, F., & Bengler, K. (2018). Are Pilots Prepared for a Cyber-Attack? A Human Factors Approach to the Experimental Evaluation of Pilots' Behavior. *Journal of Air Transport Management*, *69*, 26–37.

> »*One cannot plan for the unexpected.*«
> —*Aaron Klug*

Or maybe one can? How does an unforeseen event, as it would happen during real operations, affect pilots? In current training sessions, pilots often know what will happen. They will know which technical malfunction will arise and will prepare themselves beforehand to perform adequately in the training mission. We argue that this does not reflect reality, where problems arise suddenly and often without prior notice.

In the study presented here, we observed 22 pilots in a fixed-base simulator while they handled an unforeseeable cyberattack on their aircraft. The participants were split into two groups—control and experimental—and both groups received a warning from an air traffic controller that their aircraft might be under attack in one experimental condition. This warning introduced a high degree of uncertainty, and it also led to false alarms in the control group (as they did not actually experience an attack). We measured pilots' perceived workload, system trust, gaze behavior, and their performance regarding whether or not they made the right decisions.

The results of this study showed that a successful cyberattack has serious consequences for pilots. All dependent variables were influenced by the introduced attack: system trust was decreased, workload was increased, visual attention was directed to different instruments, and the number of wrong decisions increased by a factor of 3.3 compared to the control group. Most interesting was the fact that none of the pilots found the source of the problem; they all thought it was their mistake or attributed the abnormal behavior of the aircraft to another on-board system's malfunction.

This research was the first experimental study analyzing the effects of cyberattacks on pilot behavior. It showed that pilot performance greatly decreases when faced with totally new situations. Training should focus on giving pilots the opportunity to practice under new and unforeseen circumstances to become real experts on unpredictable situations. If training continues to avoid unforeseen events for pilots, and thus features only low content validity, it is very likely that pilots cannot transfer their trained skills to new situations.

5 Discussion on the training transfer model

In this thesis, the training transfer model developed by Baldwin and Ford (1988) was used to give an overview on different aspects influencing the training outcome and, therefore, the transfer of training. What this thesis does not do is validate whether the model is correct or whether assessment reliability (as the basis for feedback reliability) and content validity has an influence on training transfer. This was not done due to several reasons.

The first reason is that both the feedback and content validity were already found to be strong predictors of training transfer in a large number of experimental studies; see the reviews from Ford et al. (2017) and Burke and Hutchins (2007). Thus, this thesis is confined to quantifying the factors themselves. The quantification and development of appropriate metrics are seen as vital elements to sufficiently operationalize the variables of interest. Without this thesis, the lack of sufficient knowledge about the influencing factors would hinder a quantification of the relation between training transfer and those factors. So it seemed unwarranted to focus on this relation before quantifying the influencing factors.

Another reason for the decision not to validate the model was the experimental design. A classical between-group design would have been necessary to manipulate, for example, the performance assessment reliability. That is, one experimental group would experience unreliable or less reliable feedback, while another group would serve as a control group. Implementing such a design would require a considerable large number of pilots who would ideally be from the same airline and hold the same license to ensure the same preknowledge. Even then, one could argue that the previous knowledge could be different as a function of seniority, making a baseline measurement necessary. Having done so, pilots would be trained to actively fly within an airline under two different conditions, one of which might even be safety critical. Again, the control group could receive the current training while the experimental group could receive training with improved assessment and feedback reliability—but no operator would knowingly allow pilots from the same fleet to be trained differently, thereby making an experiment with such high external validity nearly impossible. Given the previous evidence from several studies, this also seemed unwarranted.

While the prediction of the individual factors is highly studied, the model does not reflect the interactions or the timely development between the influencing factors. However, it is acknowledged that those dependencies are highly context and also population specific. As previously mentioned, pilot selection plays an important role when employing pilots, but even if the selection process is highly valid, changes on the side of the pilot may arise over the years. Such changes could, for example, evolve from the work environment or the training environment. It seems reasonable that if the work environment does not feature a good transfer climate, pilots' motivation might decrease. The same is true in cases where the training design is perceived as useless. These deliberations should be considered in future research on training transfer in the aviation domain.

6 Limitations

Although limitations are discussed for every single experiment in the respective paper, some limitations might generally apply to experiments in human factors research in aviation. First, there might be constraints arising from the experimental setup. Although full flight simulators represent the highest possible degree of external validity when it comes to flight scenarios that include technical malfunctions, this does not necessarily mean that all results are transferable to the real operation. Procedures that are easily worked through in the simulator might become a real challenge in the actual operation when additional factors such as fear, time pressure, or other environmental constraints arise.

Nevertheless, the airline industry probably has no other option than to provide the highest possible external validity using the most realistic scenarios and environmental cues, such as real air traffic controller communication. In the second experiment reported in this thesis, this might be a factor influencing the results. However, as the aim was to develop and implement a new method to evaluate communication and not to measure pilot errors or how long it takes them to solve a specific problem, no significant bias is anticipated in the results obtained.

What might have biased the results is the fact that most of the pilots (all participants from Experiments #1, #2, and #3, as well as some from Experiment #4) were from the same airline. While this provides an outstanding basis to control for several aspects such as safety culture, procedural differences, and training differences, it also comes with the disadvantage that some results might not be transferable to other pilot populations. However, as we were able to randomly select participants for Experiments #1, #2, and #3, we can be sure that in terms of performance, we have a representative sample (for this specific airline).

Applying the results of this work, one has to take into account the airline-specific differences that might occur—for example, the cooperating airline provides pilots with additional training not required by the authorities, and it might have stricter selection criteria than other airlines. For the results reported here, this could mean that what was measured is better than the industry average in terms of flight performance, CRM skills, or rating skills. Airlines wanting to take advantage of this research might have to implement other or stricter countermeasures in contrast to what is proposed in the following.

7 Conclusion and recommendations

The goal of this thesis was to highlight two major factors that influence training outcomes and, thus, positive transfer of training. In the following sections, further details of the respective studies are given before deriving recommendations for the research community and airline industry.

7.1 Pilot performance assessment suffers from low interrater reliability

The results of the study presented in Paper #1 show considerable disagreement and low interrater reliability for pilots' nontechnical performance ratings. This was especially true for social aspects, such as communication. In total, we were able to show that several factors, such as the rating tool itself, familiarity with it, the scale, contextual factors, and also the rating dimension, influence the degree of within-group agreement and interrater reliability. Having chosen the most experienced raters of an airline, we can assume that the average raters would show an even smaller degree of consent.

The low interrater reliability can lead to pilots being scheduled for unnecessary extra training, causing significant training cost; however, on the other hand, pilots may be rated as performing sufficiently well when they are not (Holt et al., 2002). The biggest problem in terms of training transfer, however, is the potential lack of consistent feedback. To build up expertise, it is important to "…obtain feedback that is accurate, diagnostic, and reasonably timely" (Klein, 1999, p. 104). Because consistent feedback is a key element in building up expertise, it seems especially important for teams of high-risk reliability organizations to facilitate a training environment that features such feedback.

Recommendations:

1. Design of assessment tool: Currently, 40 items relating to four rating dimensions per pilot are used in the nontechnical skills assessment. It seems very plausible that instructors might not be able to distinguish all of them or might not have the endurance to differentiate between them when rating pilots' skills. Efforts should be taken to tighten the current assessment tool.

2. No pass/fail decisions: In line with the NOTECHS principles (Flin et al., 2003), it is important to only fail pilots based on nontechnical skill ratings in case flight safety is "...actually (or potentially) compromised. This requires a related objective technical consequence..." Flin et al. (2003, p. 109). Given the very poor reliability in the ratings on the pass/fail scale, we recommend strictly following this principle.
3. The role of the regulator: We strongly encourage regulators to explicitly advise operators to show proof of sufficient interrater reliability in their instructor population when using their rating tools. Such advice should also include specific details on how to measure reliability, what sufficient means, how raters are selected, and which scenarios have to be rated.

7.2 Cross-recurrence quantification analysis works well

In the study presented in Paper #2, we adapted and implemented a new method called cross-recurrence quantification analysis, and we used it on communication data of 12 pilot teams. While commonly used speech act theory-based measures come with very high costs in terms of coding effort, the newly developed method does not require any human coding or classification. The advantages of the new method are foremost its ability to quantify and visualize communication behavior in an objective and reliable manner.

Regarding the ability to distinguish poor from outstanding communication behavior, cross-recurrence quantification analysis shows promising results. It was found that the best crew performance is associated with dynamically changing—hence, adaptive—communication behavior that is displayed at lower recurrence rates, which is one of the metrics that can be used. Further, the new method also showed that the worst crews show significantly higher instances of simultaneous talk, which can be considered an interruption, which is an area of research that was picked up in Experiment #3.

Implementing the new cross-recurrence approach into pilot training could enhance feedback quality, as it is very transparent and directly visualizes the individuals' communication patterns. Compared to classical nontechnical skill ratings, this method does not rely on any interpretation on the side of the instructor, thus featuring the highest possible reliability. This would also free the instructor to focus more on the evaluation of other skills, which they can assess more reliably.

Recommendations:

4. Additional research needed: As with any new method, further studies should validate the results found in this experiment before implementing it into any training or assessment environment. In those studies, instructors' and pilots' opinions and their per-

ceived value of the method should be carefully considered. Further validation would also overcome the potential bias that was discussed in the limitation section regarding the population when using data from only one airline and, thus, only one population.

5. Implementation of validated algorithm: Based on the results of the first experiment on interrater reliability and on the results of the cross-recurrence quantification analyses, it seems reasonable to recommend the implementation of an objective and reliable assessment tool to measure the communication behavior of pilot crews. It is therefore recommended to use content-free approaches based on nonlinear models, such as the method implemented in this piece of research. This is foremost because linear models might not be able to correctly reflect the complexity of human interaction.

7.3 Assumptions for training and procedures are not always correct

Procedures and, hence, pilot training are based on three major assumptions: linearity, predictability, and controllability. Interruptions that occur in real operations manifest a violation of all three of those assumptions. Experiment #3 focused on whether interruptions occur during real operations. To obtain the highest possible external validity of the results, we decided to collect data from real operations in contrast to a flight simulator study. This also allowed us to investigate a phase of flight that is very seldom evaluated: the turnaround.

The results of the study showed that pilots are often interrupted and perceive a rather high workload. Although we only observed 64 errors during the 80 turnarounds, we know that interruptions come to the cost of increased error probability (Bailey & Konstan, 2006; Elfering et al., 2015; Hayes, Jackson, Davidson, & Power, 2015; Latorella, 1996; Westbrook, Woods, Rob, Dunsmuir, & Day, 2010). As aviation is such a safe system, it is very unlikely to observe any major errors leading to incidents in an experimental setup limited to 80 observations only. What is remarkable is the fact that pilots did not perceive the interruptions themselves as an issue. They often stated that it was "the way we do business here." However, this lack of awareness on the part of pilots, who do not recognize the consequences that interruptions might have, results in underestimating the risk, thus making the aviation system vulnerable.

To my knowledge, and despite the issue of interruptions being known for years, there is no dedicated training to explicitly raise awareness of the negative consequences of interruptions, nor is there training on how to handle incoming interruptions on the flight deck. That said, we observed several different strategies, showing us that pilots developed their own strategies to cope with the interruptions. For the single pilot, it might be beneficial, but the additional variance in pilot behavior might also add to a higher level of uncertainty for his/her colleague.

We believe there are several countermeasures that can be put into place to reduce the number of interruptions and their consequences. Even better, interruptions can be taken care of when future procedures and training are designed in such a way that they more accurately take the real operation into account. As interruptions are not the only perturbations that violate the assumptions made (think of, e.g., weather changes, delays, technical problems) it seems comparably easy to deal with them. However, training, as it is currently provided, seems not to represent reality sufficiently in terms of the assumptions made, thus making content validity questionable.

Recommendations:

6. Awareness training: Airlines should incorporate the topic of interruptions into their training for all personnel: pilots, flight attendants, ramp agents, and head-loaders. Such training should emphasize the negative consequences that interruptions can have and, at the same time, the important aspect of speaking up, which has to be encouraged. It is crucial to point out the need of speaking up whenever something appears to be unusual, even if this leads to interruptions.

7. Avoidance by technical means: In the medical domain, several studies have shown that very simple signs, such as "Interruptions evoke errors," can decrease the number of interruptions (Prakash et al., 2014; Relihan, O'Brien, O'Hara, & Silke, 2010). Such signs could be put on the cockpit door so approaching personnel can see them when the door is opened. Assuming that awareness training has taken place, potential interrupters would understand the risk and are likely to wait until the pilots' current task is completed before talking to them.

8. Training on handling the remaining interruptions: Pilots should receive training on how to handle situations in which they are interrupted. Depending on the situation, it might be that they write a short note or put a sticky note on the instruments that are important for upcoming tasks. Several pilots already do this, but to my knowledge there is no dedicated procedure or training that gives advice on how to handle interruptions.

9. Procedural design: Procedures should take into account that they are rarely continuously executed. Pilots are forced to interrupt them and cannot control the execution of every single item, as some items might need the input of other agents.

7.4 Unforeseen events are very critical to ensure content validity

In the last experiment we conducted, the research focus was on how pilots handle unforeseen events. In a flight simulator experiment, we presented the participants with a successful cyberattack and analyzed their behavior. Results showed an increase in workload, a decrease in system trust, a significant change in information acquisition behavior, and a decrease in the decision-making quality when under attack. This can be the basis for future research, as it was the first experiment to analyze the impacts of a cyberattack on pilots.

In the context of the research presented here, the most important finding of this experiment was that pilots experience great difficulties when they are confronted with a very new situation. When taking into account that pilots normally know what will happen during training, this piece of research shows that training must emphasize generic problem-solving and decision-making skills (see Bergström et al., 2014; Casner et al., 2013; Dahlstrom, Dekker, van Winsen, & Nyce, 2009)—some of the most important skills acquired in CRM training.

To achieve a high degree of content validity, future training programs should emphasize scenarios with uncertain situations in response to unforeseen events. Together with reliable feedback, pilots could better transfer their gathered skills in problem-solving and decision-making to real operation, for which they are not able to prepare themselves beforehand. For such training, it seems that low-fidelity simulators might even better prepare pilots for unexpected events than high-fidelity simulators (Caird, 1996; Dahlstrom et al., 2009). From our experiment and the participants' feedback, we conclude that pilots are not sufficiently prepared for completely unforeseen events such as a cyberattack.

Recommendations:

10. Awareness training: Airlines should incorporate training to facilitate cybersecurity for all personnel. Such awareness training should not only focus on pilots, but also on other personnel such as maintenance crews and flight attendants. Training has to establish a corporate and "collective awareness of cyber threats" (International Civil Aviation Organization, 2016). One approach to designing such training could be to use comparably low-cost simulation games, with the personnel receiving narratives and having to identify the root causes of the failure (e.g., Klein & Borders, 2016).

11. Cyberattack counteraction training: Further research is needed to gain knowledge on possible intrusion strategies of offenders. In accordance with these results, training has to be developed that supports pilots in identifying, understanding, and handling

infected aircraft parts. On the one hand, such training has to be generic to be applicable to several different scenarios, while on the other hand, it has to be specific enough to give hands-on advice—probably a long-term endeavor.

12. Unforeseen events in training: To increase the content validity of pilot training, there is no getting around the need to incorporate unforeseen elements. Of course, the aspects of training have to be comparable for every pilot to ensure the same level of expertise in the pilot population and to ensure fairness. One solution could be to leave the choice of some elements during training to the instructor, therefore rendering them unforeseeable to the pilots. A possible solution could be to randomly select one out of 10 scenarios at the beginning of the training mission—just enough randomness to make it not perfectly predictable for the pilots.

In light of the transfer model, opportunity to use what has been learned is an important factor that should be picked up again. As discussed earlier, pilots have very few opportunities to exercise their trained skills for emergencies. If they do, it is under very restricted conditions and always under the eyes of a trainer who assesses their success. As a final idea, I would like to recommend the following:

13. Opportunity to use: Following one of Baldwin and Ford's (1988) principles, it is recommended to create an environment in which pilots can train on their own without direct supervision. In a case where they encounter a serious incident or accident, post-hoc intervention from a trainer should be applied; in addition, pilots should have the opportunity—in addition to their normal training—to train what they think is necessary for them. As costs for full flight simulators are very high, and training transfer in abnormal situations might even be higher in low-fidelity simulators (Caird, 1996; Dahlstrom et al., 2009), costs would also be within limits.

7.5 Advancements

The results presented in this thesis focus mostly on what is not sufficiently represented in pilot training rather than on what works perfectly well. As it was the aim of this thesis to twist the knife on current issues when it comes to training transfer, it was found necessary to support the research community and airline industry with new insights and thoughts on pilot training. But it goes without saying that the aviation industry, as well as other high-risk organizations, is working on a very high level of safety. Pilots have undergone a highly valid and very restricted selection process, their training is regulated, resting periods have to follow specific rules, there are procedures and checklists for thousands of different problems, the whole of air traffic is managed by air traffic controllers monitoring the flights nearly everywhere on earth, and manufacturers constantly strive to develop the very best aircraft with the highest degree of safety possible—just to name some of the aspects that make flying so extremely safe.

One has to keep in mind that most of the aspects reported here only become dangerous in very few instances and mostly do not apply to normal operation. But those are exactly the very few instances for which we have to be prepared—even if they will never occur in a single pilot's lifetime. Besides the impact on the research community, the research presented here fortunately had several implications for the airline industry:

- More time for pilots to handle procedures: As our results showed that some procedures do not meet the needs of reality, and the assumptions underlying procedures and training are not always met in current airline operations, an attempt was started to increase the *final reserve fuel* from 30 minutes to 45 minutes (see Drees et al., 2017, for further details). Despite strong opposition, the EASA is currently evaluating the costs and benefits of increasing the *final reserve fuel*.
- Implementing unforeseen events in pilot training: In response to our recommendations to incorporate unforeseen events into pilots' recurrent training, the cooperating airline is currently adapting their training syllabus. The goal of the airline is to use some time of the four-hour shift to present pilots with unexpected and new technical and operational challenges that are not briefed before.

Although such implementations are great, one has to acknowledge that there is always room for improvement, as new threats and challenges arise every day. For example, before the introduction of the autopilot, no one thought that pilots may suffer from skill degradation of their manual flying performance; also at this time, no one thought that cyberattacks might become such a potentially dangerous topic.

8 Outlook

For the coming years, it will be a challenge not only to ensure sufficient reliability in the assessment of pilot performance, but also to ensure sufficient content validity in pilot training. To ensure appropriate pilot behavior in the very rare situations that can become dangerous, we have to build a training environment that features a realistic degree of unpredictability, so that pilots can build up experience in problem-solving and decision-making to become experts in this field of CRM. Such training has to be supported by reliable performance feedback, which is only possible when the rating tools and the rating process feature a high degree of interrater reliability.

For CRM training, implementing unforeseen events is indispensable to achieve a high level of external validity. Training paradigms might have to change to structure training to better represent reality and to rely less on assumptions that do not reflect reality. Such changes, and changes in general, come with very high costs for the airlines. It seems to be an inherent problem in aviation safety management that the advantage, cost reduction, and safety margin gained cannot easily be estimated and monetized. However, if we wait until the next serious incident or accident happens before taking corrective actions, we do not only accept potential fatalities, but will always be reactive instead of proactive.

The study on cyberattacks demonstrates the immediate need for action by the research community and the airline industry. Utilizing a proactive safety and risk management might prevent future incidents and accidents, even if we cannot exactly quantify the impact of newly developed training and countermeasures in new research areas, such as human factors in aviation cybersecurity. One important point is that the industry and the research community have to be aware of new threats and potential problems in pilot training—there is nothing worse than thinking there are no problems or demonstrating a reluctance to eliminate them. "What you don't know can't hurt you" might be true in several situations in life, but certainly not in high-risk environments such as aviation, where it is exactly the opposite: *What you don't know will hurt you!*

Contribution to individual papers

Gontar and Hoermann (2015a)

I performed the literature review, developed the research question, designed and executed the experiment, and analyzed the data. I am the main author of the paper.

Gontar, Fischer, and Bengler (2017)

I performed large parts of the literature review, developed the research question, designed and executed the experiment, and analyzed the data. I am the main author of the paper. The data was obtained within the scope of my diploma thesis where I did the analyses on speech-act distribution (Gontar, 2014), which forms the basis of comparison to the cross-recurrence quantification analyses, and is reported in Gontar, Fischer, and Bengler (2017) as well.

Gontar, Schneider, et al. (2017)

I performed large parts of the literature review, developed the initial research question, initiated the experimental design and setup, and did a complete and independent data analyses. I am the main author of the paper. The data was collected by my student within the scope of her semester thesis (Bollin, 2015). I supervised the thesis.

Gontar et al. (2018)

I performed large parts of the literature review, developed the initial research question, supported the experimental design and setup, and did a complete and independent data analyses. I am the main author of the paper. The data was collected by my students within the scope of their interdisciplinary thesis (Homans, Rostalski, Wegleiter, & Wilhelm, 2017). I supervised the thesis.

References

Abbott, K. (2010). A regulatory perspective. In B. G. Kanki, R. L. Helmreich, & J. M. Anca (Eds.), *Crew resource management* (pp. 345–360). Amsterdam and Boston: Academic Press.

Alvarez, K., Salas, E., & Garofano, C. M. (2004). An integrated model of training evaluation and effectiveness. *Human Resource Development Review*, *3*(4), 385–416. doi: 10.1177/1534484304270820

Amalberti, R., & Wioland, L. (1997). Human error in aviation. In H. M. Soekkha (Ed.), *Aviation safety* (pp. 91–108). Utrecht: VSP.

Arnold, R. (2015). Role of pilot lack of manual control proficiency in air transport aircraft accidents. *Procedia Manufacturing*, *3*, 3142–3146. doi: 10.1016/j.promfg.2015.07.862

Arora, S., Miskovic, D., Hull, L., Moorthy, K., Aggarwal, R., Johannsson, H., … Sevdalis, N. (2011). Self vs expert assessment of technical and non-technical skills in high fidelity simulation. *The American Journal of Surgery*, *202*(4), 500–506. doi: 10.1016/j.amjsurg.2011.01.024

Austin, J. L. (1962). *How to do things with words*. Cambridge, MA: Harvard University Press.

Axtell, C. M., Maitlis, S., & Yearta, S. K. (1997). Predicting immediate and longer–term transfer of training. *Personnel Review*, *26*(3), 201–213. doi: 10.1108/00483489710161413

Bailey, B. P., & Konstan, J. A. (2006). On the need for attention-aware systems: Measuring effects of interruption on task performance, error rate, and affective state. *Computers in Human Behavior*, *22*(4), 685–708. doi: 10.1016/j.chb.2005.12.009

Baldwin, T., & Ford, J. K. (1988). Transfer of training: A review and directions for future research. *Personnel Psychology*, *41*(1), 63–105. doi: 10.1111/j.1744-6570.1988.tb00632.x

Barshi, I. (2015). From healy's training principles to training specifications: The case of the comprehensive loft. *The American journal of psychology*, *128*(2), 219–227.

Bell, B. S., Tannenbaum, S. I., Ford, J. K., Noe, R. A., & Kraiger, K. (2017). 100 years of training and development research: What we know and where we should go. *The Journal of applied psychology*, *102*(3), 305–323. doi: 10.1037/apl0000142

Bengler, K., Winner, H., & Wachenfeld, W. (2017). No human – no cry? *at - Automatisierungstechnik*, *65*(7). doi: 10.1515/auto-2017-0021

Bergström, J., Dahlström, N., van Winsen, R., Lützhöft, M., Dekker, S., & Nyce, J. (2014). Rule- and role-retreat: An empirical study of procedures and resilience. *Journal of Maritime Research*, *6*(1).

Birnbach, R. A., & Longridge, T. M. (1993). The regulatory perspective. In L. Wiener, B. G. Kanki, & R. L. Helmreich (Eds.), *Cockpit resource management* (pp. 263–281). Academic Press.

Bjorklund, C., Alfredson, J., & Dekker, S. (2006). Mode monitoring and call-outs: An eye-tracking study of two-crew automated flight deck operations. *The International Journal of Aviation Psychology*, *16*(3), 263–275. doi: 10.1207/s15327108ijap1603{_}2

Blume, B. D., Ford, J. K., Baldwin, T. T., & Huang, J. L. (2010). Transfer of training: A meta-analytic review. *Journal of Management*, *36*(4), 1065–1105. doi: 10.1177/0149206309352880

Bollin, C. (2015). *Analysis of interruptions and subjective workload of pilots during turnarounds* (Semester thesis). Technical University of Munich, Munich.

Bourgeon, L., Valot, C., & Navarro, C. (2013). Communication and flexibility in aircrews facing unexpected and risky situations. *The International Journal of Aviation Psychology*, *23*(4), 289–305. doi: 10.1080/10508414.2013.833744

Brinkerhoff, R. O., & Montesino, M. U. (1995). Partnerships for training transfer: Lessons from a corporate study. *Human Resource Development Quarterly*, *6*(3), 263–274. doi: 10.1002/hrdq.3920060305

Burke, L. A., & Hutchins, H. M. (2007). Training transfer: An integrative literature review. *Human Resource Development Review*, *6*(3), 263–296. doi: 10.1177/1534484307303035

Butler, R. (1991). Lessons from cross-fleet/cross-airline observations: Evaluating the impact of crm/loft training. In R. S. Jensen (Ed.), *Proceedings of the 6th symposium of aviation psychology* (pp. 326–331). Columbus, OH: Ohio State University.

Caird, J. K. (1996). Persistent issues in the application of virtual environment systems to training. In *Third annual symposium on human interaction with complex systems* (pp. 124–132). Los Alamitos, CA: IEEE Computer Society Press. doi: 10.1109/HUICS.1996.549502

Caldwell, J. A. (2012). Crew schedules, sleep deprivation, and aviation performance. *Current Directions in Psychological Science*, *21*(2), 85–89. doi: 10.1177/0963721411435842

Caldwell, J. A., Mallis, M. M., Caldwell, J. L., Paul, M. A., Miller, J. C., & Neri, D. F. (2009). Fatigue countermeasures in aviation. *Aviation, Space, and Environmental Medicine*, *80*(1), 29–59. doi: 10.3357/ASEM.2435.2009

Carretta, T. R., & Ree, M. J. (1994). Pilot-candidate selection method: Sources of validity. *The International Journal of Aviation Psychology*, *4*(2), 103–117. doi: 10.1207/s15327108ijap0402{_}1

Carroll, J. E., & Taggart, W. R. (1987). *Cockpit resource management (CRM): A tool for improved flight safety.* Austin, TX.

Casner, S. M., Geven, R. W., & Williams, K. T. (2013). The effectiveness of airline pilot training for abnormal events. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *55*(3), 477–485. doi: 10.1177/0018720812466893

Chiaburu, D. S., & Marinova, S. V. (2005). What predicts skill transfer? an exploratory study of goal orientation, training self-efficacy and organizational supports. *International Journal of Training and Development*, *9*(2), 110–123. doi: 10.1111/j.1468-2419.2005.00225.x

Cooke, N. J., & Gorman, J. C. (2009). Interaction-based measures of cognitive systems. *Journal of Cognitive Engineering and Decision Making*, *3*(1), 27–46. doi: 10.1518/155534309X433302

Cooper, G. E., White, M. D., & Lauber, J. K. (1980). *Resource management on the flight-deck: proceedings of a nasa/industry workshop (NASA CP-2120).* San Francisco, CA.

Cooper, M. G., Elliott, E. M., & Hartzell, D. A. (1987). *U.s. patent no 4644538: Autopilot flight director system.* Washington, D.C: U.S. Patent and Trademark Office.

Cymek, D. H., Jahn, S., & Manzey, D. H. (2016). Monitoring and cross-checking automation. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *60*(1), 143–147. doi: 10.1177/1541931213601033

Dahlstrom, N., Dekker, S., van Winsen, R., & Nyce, J. (2009). Fidelity and validity of simulator training. *Theoretical Issues in Ergonomics Science*, *10*(4), 305–314. doi: 10.1080/14639220802368864

Damos, D. L. (1996). Pilot selection batteries: Shortcomings and perspectives. *The International Journal of Aviation Psychology*, *6*(2), 199–209. doi: 10.1207/s15327108ijap0602{_}6

DeChurch, L. A., & Mesmer-Magnus, J. R. (2010). The cognitive underpinnings of effective teamwork: a meta-analysis. *The Journal of Applied Psychology*, *95*(1), 32–53. doi: 10.1037/a0017328

Dedy, N. J., Szasz, P., Louridas, M., Bonrath, E. M., Husslein, H., & Grantcharov, T. P. (2015). Objective structured assessment of nontechnical skills: Reliability of a global rating scale for the in-training assessment in the operating room. *Surgery*. doi: 10.1016/j.surg.2014.12.023

Degani, A., & Wiener, E. L. (1991). Philosophy, policies, and procedures - the three p's of flight-deck operations. In R. S. Jensen (Ed.), *Proceedings of the 6th symposium of aviation psychology* (pp. 184–191). Columbus, OH: Ohio State University.

Degani, A., & Wiener, E. L. (1993). Cockpit checklists: Concepts, design, and use. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *35*(2), 345–359. doi: 10.1177/001872089303500209

Degani, A., & Wiener, E. L. (1997). Procedures in complex systems: The airline cockpit. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, *27*(3), 302–312. doi: 10.1109/3468.568739

de Zan, T., d'Amore, F., & Di Camillo, F. (2016). *The defence of civilian air traffic systems from cyber threats.* Rome: Istituto Affari Internazionali.

Drees, L., Mueller, M., Schmidt-Moll, C., Gontar, P., Zwirglmaier, K., Wang, C., … Straub, D. (2017). Risk analysis of the easa minimum fuel requirements considering the acare-defined safety target. *Journal of Air Transport Management*, *65*, 1–10. doi: 10.1016/j.jairtraman.2017.07.003

Druckman, J. N., & Kam, C. D. (2009). Students as experimental participants: A defense of the 'narrow data base'. *SSRN Electronic Journal*. doi: 10.2139/ssrn.1498843

Durso, F. T., & Alexander, A. L. (2010). Managing workload, performance, and situation awareness in aviation systems. In E. Salas & D. E. Maurino (Eds.), *Human factors in aviation* (pp. 217–247). London: Academic. doi: 10.1016/B978-0-12-374518-7.00008-0

Elfering, A., Grebner, S., & Ebener, C. (2015). Workflow interruptions, cognitive failure and near-accidents in health care. *Psychology, health & medicine*, *20*(2), 139–147. doi: 10.1080/13548506.2014.913796

Elias, B. (2015). *Protecting civil aviation from cyberattacks.*

Endsley, M. R. (1995a). Measurement of situation awareness in dynamic systems. *Human Factors*, *37*(1), 65–84. doi: 10.1518/001872095779049499

Endsley, M. R. (1995b). Toward a theory of situation awareness in dynamic systems. *Human Factors*, *37*(1), 32–64. doi: 10.1518/001872095779049543

Ephrath, A. R., & Young, L. R. (1981). Monitoring vs. man-in-the-loop detection of aircraft control failures. In J. Rasmussen & W. B. Rouse (Eds.), *Human detection and diagnosis of system failures* (pp. 143–154). Boston, MA: Springer US. doi: 10.1007/978-1-4615-9230-3{_}10

European Aviation Safety Agency. (2015a). *Annex II to ED Decision 2015/022/R.*

European Aviation Safety Agency. (2015b). *Executive Director Decision 2015/023/R: Amending Guidance Material to Part-CC of Commission Regulation (EU) No 1178/2011 AMC and GM to Part-CC - Amendment 1.*

European Commission. (2008). *Commission regulation (EC) no 859/2008: Common technical requirements and administrative procedures applicable to commercial transportation by aircraft.* Brussels.

Facteau, J., Dobbins, G. H., Russell, J. E. A., Ladd, R. T., & Kudisch, J. D. (1995). The influence of general perceptions of the training environment on pretraining motivation and perceived training transfer. *Journal of Management*, *21*(1), 1–25. doi: 10.1016/0149-2063(95)90031-4

Farrow, D. R. (2010). A regulatory perspective ii. In B. G. Kanki, R. L. Helmreich, & J. M. Anca (Eds.), *Crew resource management* (pp. 361–378). Amsterdam and Boston: Academic Press.

Federal Aviation Administration. (1996). *Part 121 - operating requirements: Domestic, flag, and supplemental operations.*

Fischer, U., McDonnell, L., & Orasanu, J. M. (2007). Linguistic correlates of team performance: Toward a tool for monitoring team functioning during space missions. *Aviation,Space, and Environment Medicine*, *78*(1), B86-B95.

Flin, R., & Martin, L. (2001). Behavioral markers for crew resource management: A review of current practice. *The International Journal of Aviation Psychology*, *11*(1), 95–118.

Flin, R., Martin, L., Goeters, K.-M., Hörmann, H.-J., Amalberti, R., Valot, C., & Nijhuis, H. (2003). Development of the notechs (non-technical skills) system for assessing pilots' crm skills. *Human Factors and Aerospace Safety*, *3*(2), 95–117.

Flynn, E. A., Barker, K. N., Gibson, J. T., Pearson, R. E., Berger, B. A., & Smith, L. A. (1999). Impact of interruptions and distractions on dispensing errors in an ambulatory care pharmacy. *American journal of health-system pharmacy : AJHP : official journal of the American Society of Health-System Pharmacists*, *56*(13), 1319–1325.

Ford, J. K., Baldwin, T. P., & Prasad, J. (2017). Transfer of training: The known and the unknown. *Annual Review of Organizational Psychology and Organizational Behavior*, *5*(1). doi: 10.1146/annurev-orgpsych-032117-104443

Ford, J. K., Quinones, M. A., Sego, D., & Sorra, J. S. (1992). Factors affecting the opportunity to perform trained tasks on the job. *Personnel Psychology*, *45*(3), 511–527. doi: 10.1111/j.1744-6570.1992.tb00858.x

Foster, J. V., Cunningham, K., Fremaux, C. M., Shah, G. H., Stewart, E. C., Rivers, R. A., … Gato, W. (2005). Dynamics modeling and simulation of large transport airplanes in upset conditions. In *Aiaa guidance, navigation, and control conference and exhibit* (pp. 15–18).

Gegenfurtner, A., Veermans, K., & Vauras, M. (2013). Effects of computer support, collaboration, and time lag on performance self-efficacy and transfer of training: A longitudinal meta-analysis. *Educational Research Review*, *8*, 75–89. doi: 10.1016/j.edurev.2012.04.001

Georgenson, D. L. (1982). The problem of transfer calls for partnership. *Training & Development Journal*, *36*(10), 75–78.

Goeters, K.-M., Maschke, P., & Klamm, A. (1998). An extended job analysis technique, the professional demands of airline pilots and implications for selection. In *Proceedings of the 23rd conference of the european association for aviation psychology.* Vienna.

Goldsmith, T. E., & Johnson, P. J. (2002). Assessing and improving evaluation of aircrew performance. *The International Journal of Aviation Psychology*, *12*(3), 223–240. doi: 10.1207/S15327108IJAP1203{_}3

Gontar, P. (2014). *Excuse me? The relation between communicational behavior and pilots' performance* (Diploma thesis). Technical University of Munich, Munich.

Gontar, P., Fischer, U., & Bengler, K. (2017). Methods to evaluate pilots' cockpit communication: Cross-recurrence analyses vs. speech act-based analyses. *Journal of Cognitive Engineering and Decision Making*, *11*(4), 337–352. doi: 10.1177/1555343417715161

Gontar, P., & Hoermann, H.-J. (2014). Flight crew performance and crm ratings based on three different perceptions. In A. Droog (Ed.), *Aviation psychology: facilitating change(s): Proceedings of the 31st eaap conference* (pp. 310–316). Malta.

Gontar, P., & Hoermann, H.-J. (2015a). Interrater reliability at the top end: Measures of pilots' nontechnical performance. *The International Journal of Aviation Psychology*, *25*(3-4), 171–190. doi: 10.1080/10508414.2015.1162636

Gontar, P., & Hoermann, H.-J. (2015b). Reliability of instructor pilots' non-technical skills ratings. In *Proceedings of the 18th international symposium on aviation psychology* (pp. 366–371). Dayton, OH: Wright State University. doi: 10.13140/RG.2.1.2152.5921

Gontar, P., Hoermann, H.-J., Deischl, J., & Haslbeck, A. (2014). How pilots assess their non-technical performance – a flight simulator study. In N. A. Stanton, S. J. Landry, G. Di Bucchianico, & A. Vallicelli (Eds.), *Advances in human aspects of transportation. part i* (pp. 119–128). AHFE International.

Gontar, P., Homans, H., Rostalski, M., Behrend, J., Dehais, F., & Bengler, K. (2018). Are pilots prepared for a cyber-attack? a human factors approach to the experimental evaluation of pilots' behavior. *Journal of Air Transport Management*, *69*, 26–37. doi: 10.1016/j.jairtraman.2018.01.004

Gontar, P., & Mulligan, J. B. (2016). Cross recurrence analysis as a measure of pilots' coordination strategy. In A. Droog, M. Schwarz, & R. Schmidt (Eds.), *Proceedings of the 32nd conference of the european association for aviation psychology* (pp. 524–544). Groningen, NL.

Gontar, P., Porstner, V., Hoermann, H.-J., & Bengler, K. (2015). Pilots' decision-making under high workload: Recognition-primed or not – an engineering point of view. In G. Lindgaard & D. Moore (Eds.), *Proceedings of the 19th triennial congress of the international ergonomics association.* Melbourne.

Gontar, P., Schneider, S. A. E., Schmidt-Moll, C., Bollin, C., & Bengler, K. (2017). Hate to interrupt you, but... analyzing turn-arounds from a cockpit perspective. *Cognition, Technology & Work*, *19*(4), 837–853. doi: 10.1007/s10111-017-0440-4

Gorman, J. C., Cooke, N. J., Amazeen, P. G., & Fouse, S. (2012). Measuring patterns in team interaction sequences using a discrete recurrence approach. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *54*(4), 503–517. doi: 10.1177/0018720811426140

Gupta, A., Li, H., & Sharda, R. (2013). Should I send this message? understanding the impact of interruptions, social hierarchy and perceived task complexity on user performance and perceived workload. *Decision Support Systems*, *55*(1), 135–145. doi: 10.1016/j.dss.2012.12.035

Haslbeck, A. (2017). *An integrative evaluation of airline pilots' manual high-precision flying skills in the age of automation* (Ph.D. thesis). Technical University of Munich, Munich.

Haslbeck, A., & Hoermann, H.-J. (2016). Flying the needles: Flight deck automation erodes fine-motor flying skills among airline pilots. *Human Factors*, *58*(4), 533–545. doi: 10.1177/0018720816640394

Haslbeck, A., Kirchner, P., Schubert, E., & Bengler, K. (2014). A flight simulator study to evaluate manual flying skills of airline pilots. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *58*(1), 11–15. doi: 10.1177/1541931214581003

Haslbeck, A., Schubert, E., Gontar, P., & Bengler, K. (2014). The relationship between pilots' manual flying skills and their visual behavior: a flight simulator study using eye tracking. In S. J. Landry, G. Salvendy, & W. Karwowski (Eds.), *Advances in human factors and ergonomics, advances in human aspects of aviation* (pp. 561–568).

Haslbeck, A., Schubert, E., Onnasch, L., Hüttig, G., Bubb, H., & Bengler, K. (2012). Manual flying skills under the influence of performance shaping factors. *Work: A Journal of Prevention, Assessment and Rehabilitation*, *41*(2012), 178–183. doi: 10.3233/WOR-2012-0153-178

Hayes, C., Jackson, D., Davidson, P. M., & Power, T. (2015). Medication errors in hospitals: a literature review of disruptions to nursing practice during medication administration. *Journal of clinical nursing*, *24*(21-22), 3063–3076. doi: 10.1111/jocn.12944

Healy, A. F., Kole, J. A., & Bourne, L. E. (2014). Training principles to advance expertise. *Frontiers in psychology*, *5*, 131. doi: 10.3389/fpsyg.2014.00131

Helmreich, R. L., & Foushee, H. C. (1993). Why crew resource management? empirical and theoretical bases of human factors training in aviation. In L. Wiener, B. G. Kanki, & R. L. Helmreich (Eds.), *Cockpit resource management.* Academic Press.

Helmreich, R. L., & Foushee, H. C. (2010). Why CRM? empirical and theoretical bases of human factors training. In B. G. Kanki, R. L. Helmreich, & J. M. Anca (Eds.), *Crew resource management* (pp. 3–57). Amsterdam and Boston: Academic Press.

Helmreich, R. L., Merritt, A. C., & Wilhelm, J. A. (1999). The evolution of crew resource management training in commercial aviation. *The International Journal of Aviation Psychology*, *9*(1), 19–32. doi: 10.1207/s15327108ijap0901{_}2

Helmreich, R. L., Sawin, L. L., & Carsrud, A. L. (1986). The honeymoon effect in job performance: Temporal increases in the predictive power of achievement motivation. *Journal of Applied Psychology*, *71*(2), 185–188. doi: 10.1037/0021-9010.71.2.185

Helmreich, R. L., Wilhelm, J. A., Kello, J. E., Taggart, W. R., & Butler, R. (1990). *Reinforcing and evaluating crew resource management: evaluator/los instructor reference manual: Technical manual 90-2.* Austin, TX.

Heppenheimer, T. A. (2003). *First flight: The wright brothers and the invention of the airplane / t.a. heppenheimer.* Hoboken, N.J. and Chichester: Wiley.

Hoermann, H.-J., Gontar, P., & Haslbeck, A. (2015). Effects of workload on sustained attention during simulator night mission. In *Proceedings of the 18th international symposium on aviation psychology* (pp. 602–607). Dayton, OH: Wright State University.

Hoermann, H.-J., & Lorenz, B. (2009). Psychologie der Luftfahrt - Forschungs- und Anwendungsgebiete. In H.-P. Krüger & N. Birbaumer (Eds.), *Enzyklopädie der Psychologie Praxisgebiete Verkehrspsychologie* (pp. 771–778). Göttingen: Hogrefe.

Holt, R. W., Hansberger, J. T., & Boehm-Davis, D. A. (2002). Improving rater calibration in aviation: A case study. *The International Journal of Aviation Psychology*, *12*(3), 305–330. doi: 10.1207/S15327108IJAP1203{_}7

Holton, E. F., Bates, R. A., & Ruona, W. E. A. (2000). Development of a generalized learning transfer system inventory. *Human Resource Development Quarterly*, *11*(4), 333.

Homans, H., Rostalski, M., Wegleiter, A., & Wilhelm, C. (2017). *Cybersecurity im Flugzeug aus ergonomischer Perspektive* (Interdisciplinary thesis). Technical University of Munich, Munich.

International Air Transport Association. (2017). *Safety report 2016* (53rd edition ed.). Montréal, Québec, Canada.

International Civil Aviation Organization. (2001). *Annex 6 to the convention on international civil aviation: Operation of aircraft: International commercial air transport - aeroplanes part i.*

International Civil Aviation Organization. (2016). *Coordinating cybersecurity work.* Montreal.

Jensen, R. S. (2017). *Presentation: great moments in isap history.* Dayton, OH.

Kanki, B. G., & Palmer, M. T. (1993). Communication and crew resource management. In L. Wiener, B. G. Kanki, & R. L. Helmreich (Eds.), *Cockpit resource management* (pp. 99–136). Academic Press.

Kanki, B. G., & Smith, G. M. (2001). Training aviation communication skills. In E. Salas, C. A. Bowers, & E. Edens (Eds.), *Improving teamwork in organizations: Applications of resource management training* (pp. 95–127). Mahwah, NJ: Lawrence Erlbaum Associates.

Karl, K. A., O'Leary-Kelly, A. M., & Martocchio, J. J. (1993). The impact of feedback and self-efficacy on performance in training. *Journal of Organizational Behavior*, *14*(4), 379–394. doi: 10.1002/job.4030140409

Kayten, P. J. (1993). The accident investigator's perspective. In L. Wiener, B. G. Kanki, & R. L. Helmreich (Eds.), *Cockpit resource management* (pp. 283–310). Academic Press.

Klein, G. (1999). *Sources of power: How people make decisions* (2nd ed.). Cambridge, MA and London: MIT press.

Klein, G., & Borders, J. (2016). The shadowbox approach to cognitive skills training. *Journal of Cognitive Engineering and Decision Making*, *10*(3), 268–280. doi: 10.1177/1555343416636515

Kontogiannis, T., & Malakis, S. (2013). Strategies in coping with complexity: Development of a behavioural marker system for air traffic controllers. *Safety Science*, *57*, 27–34. doi: 10.1016/j.ssci.2013.01.014

Kozlowski, S. W. J., Gully, S. M., Nason, E. R., & Smith, E. M. (1999). Developing adaptive teams: A theory of compilation and performance across levels and time. In D. R. Ilgen & E. D. Pulakos (Eds.), *The changing nature of performance: Implications for staffing, motivation and development* (pp. 240–292). San Francisco, CA: Jossey-Bass Publishers.

Krifka, M., Martens, S., & Schwarz, F. (2004). Linguistic factors. In R. Dietrich & T. M. Childress (Eds.), *Group interaction in high risk environments* (pp. 75–85). Aldershot: Ashgate.

Latorella, K. A. (1996). *Investigating interruptions: implications for flightdeck performance* (Unpublished doctoral dissertation). State University of New York, Buffalo, NY.

Lilienthal, O. (2003). *Der Vogelflug als Grundlage der Fliegekunst: Ein Beitrag zur Systematik der Flugtechnik* (reprint 1889 ed.). Friedland: Steffen Verlag.

Lim, D. H., & Nowell, B. (2014). Integration for training transfer: Learning, knowledge, organizational culture, and technology. In K. Schneider (Ed.), *Transfer of learning in organizations* (pp. 81–98). Cham: Springer International Publishing. doi: 10.1007/978-3-319-02093-8{_}6

Linell, P. (1982). *The written language bias in linguistics.* Linköping University Electronic Press.

Loukopoulos, L. D., Dismukes, R., & Barshi, I. (2001). Cockpit interruptions and distractions: A line observation study. In R. Jensen (Ed.), *Proceedings of the 11th international symposium on aviation psychology.* Columbus, OH.

Loukopoulos, L. D., Dismukes, R., & Barshi, I. (2003). Concurrent task demands in the cockpit: Challenges and vulnerabilities in routine flight operations. In R. S. Jensen (Ed.), *Proceedings of the 12th international symposium on aviation psychology* (pp. 737–742). Dayton, OH: Wright State University.

Loukopoulos, L. D., Dismukes, R. K., & Barshi, I. (2009). *The multitasking myth: Handling complexity in real-world operations*. Farnham: Ashgate Publishing Ltd.

Makins, N., Kirwan, B., Bettignies-Thiebaux, B., Bieder, C., Kennedy, R., Sujan, M., … Reader, T. (2016). *Keeping the aviation industry safe: safety intelligence and safety wisdom.* Eurocontrol - Future Sky Safety. doi: 10.13140/RG.2.2.14656.53763

Martinussen, M. (1996). Psychological measures as predictors of pilot performance: A meta-analysis. *The International journal of aviation psychology*, *6*(1), 1–20. doi: 10.1207/s15327108ijap0601{_}1

Marwan, N., Carmen Romano, M., Thiel, M., & Kurths, J. (2007). Recurrence plots for the analysis of complex systems. *Physics Reports*, *438*(5-6), 237–329. doi: 10.1016/j.physrep.2006.11.001

Marwan, N., & Kurths, J. (2002). Nonlinear analysis of bivariate data with cross recurrence plots. *Physics Letters A*, *302*(5-6), 299–307. doi: 10.1016/S0375-9601(02)01170-2

Mishra, A., Catchpole, K., & McCulloch, P. (2009). The oxford notechs system: reliability and validity of a tool for measuring teamwork behaviour in the operating theatre. *Quality and Safety in Health Care*, *18*(2), 104–108. doi: 10.1136/qshc.2007.024760

Mitchell, L., Flin, R., Yule, S., Mitchell, J., Coutts, K., & Youngson, G. (2012). Evaluation of the scrub practitioners' list of intraoperative non-technical skills (splints) system. *International Journal of Nursing Studies*, *49*(2), 201–211. doi: 10.1016/j.ijnurstu.2011.08.012

Mosier, K. L. (2010). The human in flight. In E. Salas & D. E. Maurino (Eds.), *Human factors in aviation* (pp. 147–173). London: Academic. doi: 10.1016/B978-0-12-374518-7.00006-7

Mosier, K. L., & Fischer, U. M. (2010). Judgment and decision making by individuals and teams: Issues, models, and applications. *Reviews of Human Factors and Ergonomics*, *6*(1), 198–256. doi: 10.1518/155723410X12849346788822

Mulligan, J. B., & Gontar, P. (2016). Measuring and modeling shared visual attention. In *Proceedings of the computational and mathematical models in vision workshop (modvis).* West Lafayette, IN.

Muthard, E. K., & Wickens, C. D. (2003). Factors that mediate flight plan monitoring and errors in plan revision: Planning under automated and high workload conditions. In R. S. Jensen (Ed.), *Proceedings of the 12th international symposium on aviation psychology.* Dayton, OH: Wright State University.

Nikandrou, I., Brinia, V., & Bereri, E. (2009). Trainee perceptions of training transfer: An empirical analysis. *Journal of European Industrial Training*, *33*(3), 255–270. doi: 10.1108/03090590910950604

Norman, D. A. (1983). Design rules based on analyses of human error. *Communications of the ACM*, *26*(4), 254–258. doi: 10.1145/2163.358092

Oakes, W. (1972). External validity and the use of real people as subjects. *American Psychologist*, *27*(10), 959–962. doi: 10.1037/h0033454

O'Connor, P., Hoermann, H.-J., Flin, R., Lodge, M., & Goeters, K.-M. (2002). Developing a method for evaluating crew resource management skills: A european perspective. *The International Journal of Aviation Psychology*, *12*(3), 263–285. doi: 10.1207/S15327108IJAP1203{_}5

Orasanu, J. M. (1993). Decision-making in the cockpit. In L. Wiener, B. G. Kanki, & R. L. Helmreich (Eds.), *Cockpit resource management* (pp. 137–168). Academic Press.

Orasanu, J. M. (1994). Shared problem models and flight crew performance. In N. Johnston, N. McDonald, & R. Fuller (Eds.), *Aviation psychology in practice.* Aldershot, Hants, England and Brookfield, Vt: Avebury Technical and Ashgate.

Orasanu, J. M., Dismukes, R. K., & Fischer, U. (1993). Decision errors in the cockpit. *Proceedings of the Human Factors and Ergonomics Society*, *37*(4), 363–367.

Orasanu, J. M., & Fischer, U. (2014). Finding decisions in natural environments: the view from the cockpit. In C. E. Zsambok & G. Klein (Eds.), *Naturalistic decision making.* Hoboken: Taylor and Francis.

Orasanu, J. M., Fischer, U., & Davison, J. (1997). Cross-cultural barriers to effective communication in aviation. In S. Oskamp & C. Granrose (Eds.), *Cross-cultural work groups: The claremont symposium on applied social psychology* (pp. 1–23). SAGE Publications.

Orasanu, J. M., Martin, L., & Davison, J. (2011). Cognitive and contextual factors in aviation accidents: decision errors. In E. Salas & G. A. Klein (Eds.), *Linking expertise and naturalistic decision making* (pp. 209–225). [Boca Raton]: Taylor & Francis.

Palmer, M. T., Lack, A. M., & Lynch, J. C. (1995). Communication conflicts of status and authority in dyadic, task-based interactions: Status generalization in airplane cockpits. *Journal of Language and Social Psychology*, *14*(1-2), 85–101. doi: 10.1177/0261927X95141005

Peterson, D. C., Martin-Gill, C., Guyette, F. X., Tobias, A. Z., McCarthy, C. E., Harrington, S. T., … Yealy, D. M. (2013). Outcomes of medical emergencies on commercial airline flights. *The New England journal of medicine*, *368*(22), 2075–2083. doi: 10.1056/NEJMoa1212052

Prakash, V., Koczmara, C., Savage, P., Trip, K., Stewart, J., McCurdie, T., … Trbovich, P. (2014). Mitigating errors caused by interruptions during medication verification and administration: interventions in a simulated ambulatory chemotherapy setting. *BMJ quality & safety*, *23*(11), 884–892. doi: 10.1136/bmjqs-2013-002484

Relihan, E., O'Brien, V., O'Hara, S., & Silke, B. (2010). The impact of a set of interventions to reduce interruptions and distractions to nurses during medication administration. *Quality & safety in health care*, *19*(5), e52. doi: 10.1136/qshc.2009.036871

Richman-Hirsch, W. L. (2001). Posttraining interventions to enhance transfer: The moderating effects of work environments. *Human Resource Development Quarterly*, *12*(2), 105–120. doi: 10.1002/hrdq.2

Rosekind, M. R., Gander, P. H., Miller, D. L., Gregory, K. B., Smith, R. M., Weldon, K. J., … Lebacqz, J. V. (1994). Fatigue in operational settings: Examples from the aviation environment. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *36*(2), 327–338. doi: 10.1177/001872089403600212

Rosenthal, R., & Rosnow, R. L. (1975). *The volunteer subject*. Oxford, England: John Wiley & Sons.

Saks, A. M. (2002). So what is a good transfer of training estimate? a reply to Fitzpatrick. *The Industrial-Organizational Psychologist*, 29–30.

Salas, E., Burke, C. S., Bowers, C. A., & Wilson, K. A. (2001). Team training in the skies: Does crew resource management (crm) training work? *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *43*(4), 641–674. doi: 10.1518/001872001775870386

Salas, E., Cooke, N. J., & Rosen, M. A. (2008). On teams, teamwork, and team performance: Discoveries and developments. *Human Factors*, *50*(3), 540–547. doi: 10.1518/001872008X288457

Salas, E., Sims, D. E., & Burke, C. S. (2005). Is there a ”big five” in teamwork? *Small Group Research*, *36*(5), 555–599. doi: 10.1177/1046496405277134

Sarter, N. B., Mumaw, R. J., & Wickens, C. D. (2007). Pilots' monitoring strategies and performance on automated flight decks: An empirical study combining behavioral and eye-tracking data. *Human Factors*, *49*(3), 347–357. doi: 10.1518/001872007X196685

Sarter, N. B., & Woods, D. D. (1992). Pilot interaction with cockpit automation: Operational experiences with the flight management system. *The International journal of aviation psychology*, *2*(4), 303–321. doi: 10.1207/s15327108ijap0204{_}5

Sarter, N. B., & Woods, D. D. (1994). Pilot interaction with cockpit automation II: An experimental study of pilots' model and awareness of the flight management system. *The International journal of aviation psychology*, *4*(1), 1–28. doi: 10.1207/s15327108ijap0401{_}1

Schindler, L. A., & Burkholder, G. J. (2016). A mixed methods examination of the influence of dimensions of support on training transfer. *Journal of Mixed Methods Research*, *10*(3), 292–310. doi: 10.1177/1558689814557132

Seamster, T., Hamman, W., & Edens, E. (1995). Specification of observable behaviors within loe/loft event sets. In R. S. Jensen (Ed.), *Proceedings of the 8th symposium of aviation psychology* (pp. 663–668). Columbus, OH: Ohio State University.

Searle, J. R. (1969). *Speech acts: An essay in the philosophy of language* (Reprinted ed.). Cambridge [u.a.]: Cambridge Univ. Press.

Sears, D. O. (1986). College sophomores in the laboratory: Influences of a narrow data base on social psychology's view of human nature. *Journal of Personality and Social Psychology*, *51*(3), 515–530. doi: 10.1037/0022-3514.51.3.515

Sevdalis, N., Davis, R., Koutantji, M., Undre, S., Darzi, A., & Vincent, C. A. (2008). Reliability of a revised notechs scale for use in surgical teams. *The American Journal of Surgery*, *196*(2), 184–190. doi: 10.1016/j.amjsurg.2007.08.070

Shappell, S. A., & Wiegmann, D. A. (1997). A human error approach to accident investigation: The taxonomy of unsafe operations. *International Journal of Aviation Psychology*, *7*(4), 269–291. doi: 10.1207/s15327108ijap0704{_}2

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86*(2), 420–428.

Smith, H. P. R. (1979). *A simulator study of the interaction of pilot workload with errors, vigilance, and decisions.* Moffett Field, CA.

Straus, S. G., & Cooper, R. S. (1989). Crew structure, automation and communication: Interaction of social and technological factors on complex systems performance. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *33*(13), 783–787. doi: 10.1177/154193128903301303

Thomas, L. C., & Rantanen, E. M. (2006). Human factors issues in implementation of advanced aviation technologies: A case of false alerts and cockpit displays of traffic information. *Theoretical Issues in Ergonomics Science*, *7*(5), 501–523. doi: 10.1080/14639220500090083

Turner, M., Griffin, M. J., & Holland, I. (2000). Airsickness and aircraft motion during short-haul flights. *Aviation, space, and environmental medicine*, *71*(12), 1181–1189.

U.S. Department of Transportation. (2004). *Advisory circular: Crew resource management training: Ac no. 120-51e.* Landover, MD.

Vidulich, M. A., Wickens, C. D., Tsang, P. S., & Flach, J. M. (2010). Information processing in aviation. In E. Salas & D. E. Maurino (Eds.), *Human factors in aviation* (pp. 175–215). London: Academic. doi: 10.1016/B978-0-12-374518-7.00007-9

Webber, C. L., & Marwan, N. (Eds.). (2015). *Recurrence quantification analysis*. Cham: Springer International Publishing. doi: 10.1007/978-3-319-07155-8

Weber, D. E., Mavin, T. J., Roth, W. M., Henriqson, E., & Dekker, S. W. A. (2014). Exploring the use of categories in the assessment of airline pilots' performance as a potential source of examiners' disagreement. *Journal of Cognitive Engineering and Decision Making*, *8*(3), 248–264. doi: 10.1177/1555343414532813

Weigl, M., Antoniadis, S., Chiapponi, C., Bruns, C., & Sevdalis, N. (2015). The impact of intra-operative interruptions on surgeons' perceived workload: an observational study in elective general and orthopedic surgery. *Surgical endoscopy*, *29*(1), 145–153. doi: 10.1007/s00464-014-3668-6

Weigl, M., Müller, A., Vincent, C., Angerer, P., & Sevdalis, N. (2012). The association of workflow interruptions and hospital doctors' workload: a prospective observational study. *BMJ quality & safety*, *21*(5), 399–407. doi: 10.1136/bmjqs-2011-000188

Weiland, M., Nesthus, T., Compatore, C., Popkin, S., Mangie, J., Thomas, L. C., & Flynn-Evans, E. (2013). Aviation fatigue. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *57*(1), 1–5. doi: 10.1177/1541931213571001

Weissbein, D. A., Huang, J. L., Ford, J. K., & Schmidt, A. M. (2011). Influencing learning states to enhance trainee motivation and improve training transfer. *Journal of Business and Psychology*, *26*(4), 423–435. doi: 10.1007/s10869-010-9198-x

Westbrook, J. I., Woods, A., Rob, M. I., Dunsmuir, W. T. M., & Day, R. O. (2010). Association of interruptions with an increased risk and severity of medication administration errors. *Archives of internal medicine*, *170*(8), 683–690. doi: 10.1001/archinternmed.2010.65

Wilshusen, G. C. (2013). *Cybersecurity: A better defined and implemented national strategy is needed to address persistent challenges.* Washington, D.C..

Yule, S., Flin, R., Maran, N., Rowley, D., Youngson, G., & Paterson-Brown, S. (2008).

Surgeons' non-technical skills in the operating room: Reliability testing of the notss behavior rating system. *World Journal of Surgery*, *32*(4), 548–556. doi: 10.1007/s00268-007-9320-z

Yule, S., Rowley, D., Flin, R., Maran, N., Youngson, G., Duncan, J., & Paterson-Brown, S. (2009). Experience matters: comparing novice and expert ratings of non-technical skills using the notss system. *ANZ Journal of Surgery*, *79*(3), 154–160. doi: 10.1111/j.1445-2197.2008.04833.x

Zijlstra, F. R. H., Roe, R. A., Leonora, A. B., & Krediet, I. (1999). Temporal factors in mental work: Effects of interrupted activities. *Journal of Occupational and Organizational Psychology*, *72*(2), 163–185. doi: 10.1348/096317999166581